A Turing Test for Computer Game Bots

Philip Hingston, *Senior Member, IEEE*

***Abstract*—In this paper, a version of the Turing Test is proposed, to test the ability of computer game playing agents ("bots") to imitate human game players. The proposed test has been implemented as a bot design and programming competition, the 2K BotPrize Contest. The results of the 2008 competition are presented and analyzed. We find that the Test is challenging, but that current techniques show promise. We also suggest probable future directions for developing improved bots.**

***Index Terms*—Artificial intelligence, competitions, games, intelligent systems.**

I. Introduction

FOR more than half a century, the Turing Test has been a challenge and an inspiration for artificial intelligence researchers. It has also been the subject of much philosophical debate, and continues to be a touchstone for discussions of what it means for a computing machine to be intelligent.

The Turing Test was first proposed in Alan Turing's 1950 paper [1], in which Turing considers the question "Can a Machine Think?" He rejects argument based on semantic word play, labeling the question meaningless, proposing instead an objective test that we now know as the Turing Test. The Test is commonly described something like this (although this is not exactly as Turing described it):

> Suppose you are chatting with some entity via a text chat session. Could you tell, solely from the conduct of the session, whether the other entity was a human or a machine? If not, then that entity is judged to be intelligent.

There has been much debate about whether this test is a valid test for intelligence, or whether it tests something else. Our view, similar to that of many others, is that it is a test of the *ability to appear to be human*, and that being intelligent is one possible way to do that. (Weizenbaum's ELIZA program, for example, used simple textual pattern matching to imitate the responses of a Rogerian psychotherapist.)

There is an important new class of computer applications in which the ability to appear to be human is of key importance—interactive video games. Computer games, in general, have always had a strong association with artificial intelligence, and modern computer games are seen as a rich application area and testbed for artificial intelligence research. In particular, interactive video games often present the human player with a virtual world, populated with characters, usually humanoid, that may be controlled either by other human players or by software. Software-controlled characters are sometimes called "bots." The commercial success of the game can hinge, to a large extent, on how convincingly these bots are able to impersonate a human player.

With this in mind, we propose a variation of the Turing Test, designed to test the abilities of computer game bots to impersonate a human player. The Turing Test for (computer game) bots is as follows:

> Suppose you are playing an interactive video game with some entity. Could you tell, solely from the conduct of the game, whether the other entity was a human player or a bot? If not, then the bot is deemed to have passed the test.

At the 2008 IEEE Symposium on Computational Intelligence and Games (CIG) in December 2008, the author ran a competition, known as the 2K BotPrize Contest, challenging entrants to create a bot that could pass this test. If no one could pass the test, a secondary aim was to see who could create the most humanlike bot, and to learn what we could from the event. Games development company 2K Australia (Canberra, A.C.T., Australia) provided a cash prize of A$7000 for anyone who could pass the test, or A$2000 for the most humanlike bot. In this paper, we report on the competition and its outcomes. Although no one passed the test, some useful lessons were learned.

The structure of this paper is as follows. In the next section, we provide background and motivation for proposing a Turing Test for Bots. In Section III, we review related work, before introducing our proposed test in Section IV. The following section offers a discussion of the philosophical implications of the test, based on the philosophical literature discussing the Turing Test. Section VI describes the BotPrize Contest. The results of the competition are presented in summarized form in Section VII, and analyzed and discussed in the following section. Plans for future research and competitions are outlined in Section IX. The final section summarizes our findings. Appendixes I and II provide the raw data collected during the competition.

II. Background

It is a curious fact that computers and games seem to have a natural connection. From the earliest days of computing, games have been used as example domains to test the thinking abilities of computers, while in modern times, a major application of computers is to provide entertainment through games.

This Turing Test for Bots has its roots in both arenas. On the one hand, it is obviously inspired by a game that Turing first proposed, as a way to answer questions about computers and thinking. On the other, it is also about how artificial intelligence/computational intelligence (AI/CI) can be used to improve modern computer games.

Manuscript received June 15, 2009; revised August 18, 2009; accepted September 05, 2009. First published September 18, 2009; current version published December 01, 2009. The BotPrize competitions, both in 2008 (the inaugural year) and in 2009, were supported by 2K Australia.

The author is with Edith Cowan University, Mt. Lawley, Perth, 6050 W.A., Australia (e-mail: p.hingston@ecu.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCIAIG.2009.2032534

There are two reasons for proposing the Turing Test for Bots and organizing the 2K BotPrize Contest: to challenge AI/CI researchers to advance the state of the art in AI/CI as applied to modern computer games, and to contribute to the study of machine intelligence.

Why is it necessary to advance the state of the art in AI/CI for computer games? There are many potential applications for AI/CI in modern computer games, i.e., interactive video games, yet their use is not as prevalent as academic researchers would like it to be. Why is this? We suggest that one important reason is that game developers believe that sophisticated AI/CI is not required for a successful game.

Recently, when the author approached a games development entrepreneur to try to interest him in collaboration, his answer was "but Game AI is a solved problem." A deeper discussion did uncover some interesting areas for research and development, but that initial response is often as far as the conversation goes. Curiously, a colleague recently reported that when he approached someone in the industry about collaboration, he got the reply "but Game AI is too hard." We hope that the BotPrize Contest will be a demonstration and a launch pad for new techniques that will be valued and used by game developers.

At the 2008 CIG, one of the keynote presentations [2] was by Jason Hutcheons, Lead Programmer at Interzone Games (Jason is also a past winner of the bronze medal in the Loebner Prize Contest). One of his main themes was that game developers are happy to "cheat" to get the effects they want in their games and that if something in a game has to look intelligent, then the *appearance of intelligence* is all that is needed. He used the term pseudointelligence (borrowed from *The Diamond Age*, Neal Stephenson's novel based on the Turing Test), as a better term than artificial intelligence to emphasize this distinction.[1]

Some of the well-known applications for AI/CI in video games are real-time decision making (e.g., planning and path finding), and giving other behavior to artificial characters in games, whether they be opponents, teammates, or incidental characters. These artificial players are sometimes called nonplayer characters (NPCs), or AIs, or bots. There are also less known applications such as story generation, data mining of game logs, optimizing game design, and automating the movement of virtual cameras.

As a challenge problem, we decided to focus on bots. One reason for this choice is that bots are a very visible game component that most game players would associate with artificial intelligence. Another is that, at least anecdotally, a poorly implemented bot is an indicator of a poor quality game, and reduces the enjoyment (and the popularity and commercial success) of the game.

An example of a poorly implemented bot would be one that plays at a superhuman level, and so is frustrating to play against. Another example would be a bot whose behavior is so predictable that it can be exploited in some way, whereas a human would realize what was happening and change behavior. In both cases, the problem is that the bot does not appear to be human.

[1]Jason's presentation is available as an audio/slide recording on the web at http://www.csse.uwa.edu.au/cig08/JasonHutchens.html and we recommend it for researchers who would like to work with commercial game companies.

This anecdotal evidence, supported by recent experimental results [12] that humanlike bots are more fun to play, suggests that a good way to test the quality of a bot would be to test its ability to appear human. From now on, we will use the term *humanness* to mean the ability to appear human.

For these reasons, we believe that the problem of making a bot appear human is an apt one: it addresses a problem with commercial significance, and is *a priori* a task where AI/CI techniques should be highly relevant. We would expect that core machine intelligence capabilities such as adaptation and learning, planning, and opponent modeling, among others, would be needed to make a bot appear human. But this is an expectation that needs to be tested.

How would one go about testing for humanness? One way is to implement a test similar to the Turing Test. In the next section, we introduce the original Turing Test, and discuss other related work on making bots appear human, before going on to propose a test for humanness inspired by the Turing Test.

III. Related Work

In this section, we introduce the original Turing Test, which is the model on which our proposed test is based. We then discuss the Loebner Prize, a competition based on the Turing Test, just as the BotPrize Contest is based on our Turing Test for Bots. We also review other existing work on bot humanness.

A. The Turing Test

The Turing Test was originally described by Turing in terms of a game—the "imitation game"[1]. The imitation game is a very curious game, and we hazard that most people who know of the Turing Test are not aware of its original form.

In this game, there are three participants: an interrogator, a woman, and a competitor (who is a man). The interrogator knows the other two as X and Y, and his task is to determine which of them is the woman. To do this, the interrogator asks them questions like "X, how long is your hair?" Questions and answers are exchanged using a teletype machine (the equivalent of a modern chat session). The woman is assumed to be trying to help the interrogator with her answers, while the man's task is to deceive the interrogator. Could the man deceive the interrogator 10% of the time or 20% of the time or 50% of the time? Turing proposed this game of deceit as a test of the man's thinking ability.

Turing's contention was that, if a computer could play the imitation game as well as a man, then the computer must be, in some sense, thinking (i.e., it must be an intelligent entity).

Later in [1, p. 442], the woman disappears and the imitation game is recast as a game with three participants: an interrogator, a man, and a computer. The interrogator now has to identify the man. The man tries to assist the interrogator, while the computer tries to deceive them. Turing then makes his famous prediction that by the year 2000, a computer would be able to deceive the interrogator at least 30% of the time, given five minutes of questioning. Note that, in modern versions, gender is removed entirely from the test, replacing "man" above with "human."

B. The Loebner Prize

Since 1991, Hugh Loebner has been organizing an annual competition, known as the Loebner Prize, based on the Turing Test. While the exact conditions have changed slightly over the years, it is in essence a competition for programmers to create a program (a "chatterbot" or "chatbot") that can hold a chat session with an interrogator, during which session the interrogator tries to determine whether they are chatting with a human or a program. The term "judge" is used to mean an interrogator, and "confederate" is used to refer to the human in the imitation game. We will use the same terminology from now on.

The format for the 2009 competition is as follows [3].

- Each game is between a judge, a chatbot, and a confederate.
- The judge converses with the chatbot and confederate using a split-screen chat program. There is no interaction between the chatbot and the confederate, and neither can see the judge's conversation with the other. Thus, the conversations are essentially independent.
- The interface is such that the judge can watch the others "typing" their answers—mistakes can be corrected and characters appear in "real time." (This introduces a behavioral aspect to the test that was not present in the Turing Test.)
- After 10 min of conversation (in previous years the time limit was 5 min), the judge has 10 min to consider, and then must nominate one or the other as the human.
- There is no restriction on the topics of conversation (in early years of the contest, the topics were restricted).
- There are four judges, four confederates, and four chatbots in the final. (This has varied; for example, in 2008, there were 12 judges.)
- Every judge meets every chatbot and every confederate once, but not in the same combinations.
- After all the games, each judge rates the four entities they judged to be nonhuman on a humanness scale from 1 to 4. (The average rating is to be used in case of a tie.)

No chatbot has ever passed the test (the first to do so wins a US$100 000 prize), but, for example, the 2008 winner, Elbot, deceived three out of 12 judges. A number of commercially successful chatbots, including Elbot, had their origins as Loebner Prize entrants.

C. Bots Imitating Humans

Researchers have used a variety of techniques to try to make their bots more humanlike. A few examples are as follows: behavior-based techniques from robotics [5] were applied in [6]; Tatai and Gudwin [7] introduced a technique called "semionics" to control bot actions, with the aim of creating humanlike bots; and hidden semi-Markov models and particle filters were used in [8] to predict opponent movement, and the performance of the resulting bots was compared with human performance. Some researchers have not only created bots based on AI/CI methods, but have also proposed tests to measure the humanness of their bots.

In [9], Laird and Duchi examined some factors that make bots appear more humanlike. One reason for their interest in humanness of bots was their work on simulating human pilots using Soar [10]. They made modifications to bots that played the first-person shooter Quake. Although their results were only indicative (due to high variance in the data), they found that decision time and aiming skill were critical factors in making bots seem more humanlike, whereas aggressiveness and number of tactics were not. Their methodology for evaluating humanness was a kind of Turing Test, except that the judges did not actually interact with the bots. Video recordings were made of games played from the point of view of the bot. Similar recordings were made from the point of view of five human players of different skill levels. A panel of eight judges, of varying skill levels, viewed the videotapes (three judges viewed each tape) and made judgments about the humanness and skill levels of the players (human and bot). The judges were asked to assign humanness ratings, and also to make a binary decision as to whether the player was human. They achieved an accuracy of 16/18 humans correctly identified, and 27/48 bots correctly identified.

In [11], Gorman *et al.* used imitation learning (in this case, reinforcement learning) to "train" bots in navigation and movement tasks, and then tested their humanness using a "believability test." In this test, subjects are shown video clips of bot behaviors and asked about their game playing expertise, and then asked to rate the likelihood that the bot is human on a scale from 1 to 5. These data were then combined to calculate a kind of weighted average called the "believability index." They found that, according to this test, their bots were just as believable as humans.

In [12], Soni and Hingston report on experiments in which subjects played games against standard scripted bots, and against bots trained to play "like a human" (the bots used a neural network, trained on recorded examples of human play). Subjects were then asked which bots they preferred to play, which were more challenging as opponents, which were predictable, and which appeared more humanlike. Subjects reported the trained bots as being more humanlike and more fun to play against.

IV. A Turing Test for Bots

One might say that the primary focus of the Turing Test is machine intelligence, and that the task of imitating a human is simply a convenient, concrete task for demonstrating it. Contrariwise, our Turing Test for Bots is primarily about imitating a human, and machine intelligence is one possible way to do the imitating.

The test is based on an imitation game between an interrogator (judge), a human (confederate), and a computer program (bot). The three participants play a video game with each other. The judge tries to identify the human based on the behaviors of the other two players. The bot tries to deceive the judge into thinking it is the human (or rather the bot's programmers try to program the bot so as to deceive the judge).

Should a chatbot pass the Turing Test, this would presumably be taken by some as evidence that the chatbot is intelligent. In contrast, should a game playing bot pass the Turing Test for Bots, we would make no such claim.

While this test is superficially similar to the Turing Test, there are a number of differences, some subtle.

Fig. 1. *Unreal Tournament 2004* screenshot showing the first-person 3-D view of the game world.

- The three participants are taking part in a three-way interaction, not in paired two-way interactions as in the Turing Test and the Loebner Prize.
- The confederate is not trying to assist the judge, but instead is indifferent to the imitation game. The confederate simply wants to win the video game.
- The task is much more restricted than that of carrying out a conversation in natural language.
- The underlying representation supplied to humans and bots is not identical: for humans, it is vision and sound, and for bots it is data and events. This is discussed in detail in Section IV-A.

A. Underlying Representation

In the same way that the Turing Test has been operationalized in the form of the Loebner Prize Contest, this Turing Test for Bots has been operationalized in the form of the 2K BotPrize Contest. But there are some differences in the way the two contests are run. We describe the design of the 2008 2K BotPrize Contest in detail in Section VI.

In the Turing Test, the participants interact with and perceive each other only via text messages. In the Turing Test for Bots, humans and bots necessarily perceive the virtual world (including each other) via different underlying representations.

For the contest, we chose to use a commercial computer game called *Unreal Tournament 2004* (UT2004). UT2004 is what is known as a "first-person shooter" (FPS). The player uses a keyboard and mouse to control the actions of a character in a 3-D virtual world, and the screen shows a 3-D view of the world from the point of view of the player. The player moves his character around the virtual world, usually with the aim of doing simulated violence to other characters in the game (e.g., by shooting them). Doing enough damage results in simulated death for the other character (this is known as a "frag"). Being fragged is usually only a temporary setback—the character is returned to the game in full health ("respawned") after a short delay.

For humans, the primary input is a real-time 3-D first-person view, along with sound effects. The player and opponents are represented by humanoid avatars. Fig. 1 shows a typical view. In this screenshot, you can see an opponent (player 120), a "pickup" (where items such as weapons and health packs appear or "spawn" from time to time—in this case, a weapon has spawned), and the layout of a section of the map including a door behind the opponent, and a ramp down to another level just to the right. There is also information such as the current score, what weapons the player is carrying and currently using, the player's current health level and so on, presented as a heads-up display.

The range of actions available to a character in the game include walking forwards, backwards, and sideways, running, turning, crouching, jumping, picking things up, dropping things, operating weapons, throwing projectiles, as well as verbal communications like taunting opponents and giving instructions to teammates, and special moves (such as turning invisible) known as "combos."

For bots in *Unreal Tournament 2004*, an interface known as GameBots provides input data in the form of messages. GameBots clients receive information about the game (for example, the locations of objects, or information about other players) and send commands to control their bots. Information can be received in response to a request from the client (for example, clients can ask what items their bot is carrying, or can cast a ray into the virtual world and get back information about what the ray hits), or as a result of events in the game (for example, another player becoming visible, a sound being heard, or the bot being damaged). There are control commands for such things as changing weapons, walking, strafing, shooting, jumping, and so on. These facilities give the bot similar information and possible actions to those that a human player would have. A complete listing of the GameBots application programming inter-

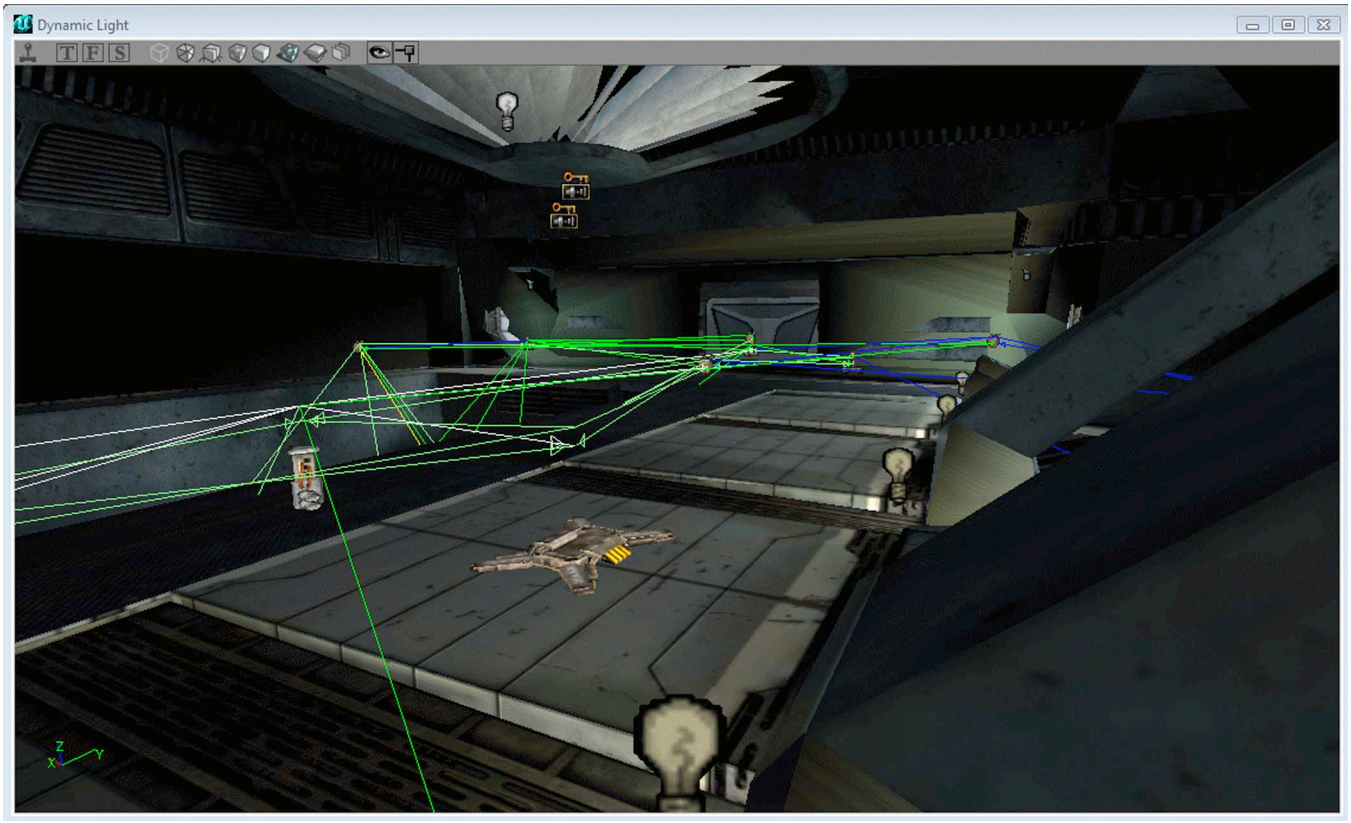

Fig. 2. *Unreal Tournament 2004* map showing navigational mesh.

face (API) can be found in [4]. Here is a summary of the most relevant messages.

- At the start of a round, information describing a network of map locations (known as navpoints) is sent to the bot, including for each point:
  - location (x, y, and z coordinates);
  - what items spawn here;
  - pointers to neighboring navpoints;
  - information on what movements are possible to get to and from this location;
  - whether this is a good ambush point;
  - whether this is a good sniping point;
  - what weapon should be used from here.
- Also at the start of the round, information about all pickup items on the map is sent to the bot, including for each item:
  - location (x, y, and z coordinates);
  - item type.
- Messages are sent to a bot during the game when various events occur, including:
  - adrenaline gained;
  - weapon or ammunition picked up;
  - bumped by another player or object;
  - current weapon changed;
  - damage taken;
  - bot has died;
  - bot has just hit a ledge;
  - bot is falling down;
  - hit another player;
  - bot hears a noise;
  - another player died;
  - bot has lost some item;
  - a projectile is incoming;
  - another player is seen (messages generated 1 or 2 times per second);
  - bot just collided with a wall.

These event messages contain information such as who instigated damage taken (if the other player is visible), who killed the bot, who killed the other player, how much damage was taken, etc.

The bot can also request certain information.

- Check whether it is possible to move to a given location from the current location.
- Send a "ray" into the world and find out what it hits (the ray has no effect in the game; this just provides information).
- Get a path from one location to another.

In Fig. 2, we see a view of the same screen as in Fig. 1, showing part of the network of nodes, known as the navigational mesh.

Most of the above information would be available to a human player via vision and audio, although some would have to be learned (for example, by building up a navigational mesh while moving around the map).

V. Philosophical Implications of the Turing Test for Bots

In this section, we briefly outline a few of the many responses to Turing's proposed test and relate them to our proposed test for bots. In doing so, we do not attempt to enter into debate on

the merits of the Turing Test; our intent is simply to reflect on the proposed Test for bots in the light of these responses. The exposition of these responses is influenced by Saygin *et al.*'s review of the Turing Test [13], though we have given our own interpretations and omitted parts that do not seem relevant for our purposes. We refer the interested reader to that paper for more detailed and complete discussion and references.

The value of the Loebner Prize Contests has also been debated, sometimes acrimoniously. We will not enter into this debate either, but refer the interested reader to one example: Schieber's critique of the contest [14] and Loebner's reply [15]. Responses to the Turing Test can be divided into broad categories.

A. Objections That Do Not Apply

Many objections to the Turing Test do not apply to the Turing Test for Bots, because the Test for Bots is intended only to test whether a bot has the appearance of being human, not whether it is actually human, whereas the Turing test is claimed to test whether the machine is *actually thinking*, not just appearing to be thinking. In our case, we are agnostic about whether, in appearing to be human, our bots are actually thinking (or self-aware, or conscious), and it is obvious that bots are not actually human.

Purtill [16], for example, argues that if a machine were good at the imitation game, then that would indicate thinking on the part of its programmer, but not on the part of the machine. In the Turing Test for Bots, we admit *a priori* that the bot is not human.

B. Objections That Are Out of Scope

There are other objections that may or may not apply to the Turing Test for Bots, but even if they did, they would not reduce its value for our purposes. By and large, these are complex philosophical arguments about the true nature and meanings of things. While these may be valid and interesting questions for philosophers, they are not our main concern here.

An example is Gunderson's argument [17] that a machine might be able to give the appearance of a thing, without imitating it. He proposes a test that he says is similar to the original imitation game, the "toe stepping game," where the contestant attempts to convince a judge that he/it is a woman, stepping on their toe. He then claims we could build a machine to play the game using a box containing rocks and a trigger mechanism, but that such a machine would not be imitating a woman in the same sense that a man would be.

As far as the Turing Test for Bots is concerned, whether a bot is actually imitating a human or just giving the appearance of imitating a human (if this distinction is meaningful) is not critical. It is the end result that interests us.

Incidentally, as Dennett [18], for example, has noted, the toe-stepping argument is the sort of semantic argument that Turing was trying to avoid by proposing his test.

C. The Test Is Too Narrow

This argument is exemplified by Gunderson's "all-purpose vacuum cleaner" objection:

> A vacuum cleaner salesman claims his product is all-purpose, and demonstrates by vacuuming up some dust. When asked what else it does–can it suck up straw? Or rocks?–he answers that no, vacuum cleaners are for sucking up dust. Is he justified in claiming his product is all-purpose?

Gunderson [17] claims that this is analogous to the imitation game; a machine that can play the imitation game shows that it can do one thing: hold a conversation, but an intelligent being should be able to do many other things as well. The usual counterargument is that a good player of the imitation game would need many abilities (reasoning, calculation, memory, etc.).

The argument could also be applied to the Turing Test for Bots, and we would make a similar reply. Although the test is not as challenging as the imitation game, doing a good job of imitating a human player requires many abilities (navigation, selection of appropriate weapons, tracking opponents, dodging, hiding, etc.).

D. The Test Is Anthropomorphic

Millar [19] objects that the Turing Test is a test of *human* intelligence, rather than a test of intelligence *per se*. We think this argument is mistaken (just because the machine has to imitate human intelligence does not mean it has to use human intelligence to do the imitating).

But as far as the Turing Test for Bots goes, it is not necessary to counter this argument. The Turing Test for Bots is intentionally anthropomorphic.

E. Psychologism and Behaviorism

Block [20] argues that the Turing Test is inadequate because it does not consider the "character of the internal information processing" in the machine. Psychologism is the position that whether a machine really thinks depends on this, behaviorism is the position that it does not. The typical counterargument is that psychologism is chauvinistic: imagine an intelligent Martian life form that uses different internal information processing than humans use—could it not still be intelligent?

This objection might be included with objections that are out of scope, but there is an interesting parallel here between psychologism versus behaviorism and pseudo- versus artificial intelligence that is worth noting. As far as the Turing Test for Bots goes, it does not matter what the internal mechanisms of the bot are; there is no concept of "cheating" and the only constraint on the internal mechanism is that it has to work in a practical sense. Of course, as AI/CI researchers, we are keenly interested in these internal mechanisms.

We would put Block's "Aunt Bubbles machine" argument [21] and Searl's "Chinese Room" [22] into the first category of objections. They do not apply to the Test for bots.

F. Human Traits Unrelated to Intelligence

Here we group together a family of objections to the Turing Test based on aspects of human psychology/physiology that would be hard to imitate, but are accidental phenomena rather than hallmarks of intelligence.

Michie [23] gives the example of pluralizing nonsense words: "How do you pronounce the plurals of the imaginary English words "platch," "snorp," and "brell"?" might be answered with "I would pronounce them as "platchez," "snorpss," and "brellz."" A human would do this using some subconscious process, the argument goes, but it would be unreasonably difficult for a machine to anticipate this question and be able to simulate the process, say with a suitable set of rules.

French [24] gives other examples based on "subcognition." Here is one: for humans, the time required to recognize the second word of a pair depends on whether the first word is a related word, an unrelated word, or a nonword. Again, it might be possible for a machine to calculate and exhibit the right kinds of recognition delays, but this might be considered unreasonably difficult.

These examples would not refute the validity of the Turing Test for Bots, which is a test of humanness, because they are in fact valid examples of behaviors expected of humans. But they do bring to our attention something we should be concerned with: if there are equivalent phenomena that would affect a human judge's perception of a player's humanness, then a good bot should imitate those phenomena.

Having introduced and discussed the Test for Bots and its origins, we will now describe how it was carried out in the form of a competition.

VI. The Competition

The Loebner Prize competitions exemplify one way to try to answer Turing's question, "Can a machine think?" What kind of experiment could we carry out in order to answer *our* question?

We followed Loebner's example, and proposed a competition that challenges entrants to program an agent (in our case, a computer game bot) that can fool a panel of judges. The competition was run as part of the 2008 IEEE Symposium on Computational Intelligence and Games, in December 2008.

The detailed design of the competitions was as follows.

A. Choice of Game

As mentioned earlier, the competition makes use of the commercial computer game *Unreal Tournament 2004*. There were a number of reasons for this choice.

- It is well known among game players.
- It is readily available and inexpensive.
- It is relatively easy to interface a bot to.
- It provides for multiple human players and bots to play together.
- The game format is suitable for a number of short rounds of competition.
- It is possible to customize the game to a degree.

The game play takes place in a virtual world, simulated by a program known as the server. The player connects to the server over a network, using another program known as a client. Other characters can be controlled by other players via their own clients. The game also allows for characters controlled by inbuilt bots running on the server.

In addition, the game has its own scripting language (UnrealScript), which allows hobbyists and developers (known as "modders") to modify certain aspects of the game via hooks provided by the game's designers. Several teams of modders have used this facility to create a modification to the game (a "mod") that allows characters in the game to be controlled and to receive game information via network sockets. This allows programmers to write external programs to control bots in the game.

One such group is the Pogamut team.[2] Pogamut [25] uses a version of the GameBots mod to interface with UT2004, and provides an integrated development environment (IDE) for programmers, integrated with NetBeans, a Java IDE. GameBots originated as a mod for *Unreal Tournament*, created at the University of Southern California's Information Sciences Institute, Los Angeles, in 2000, and since then a number of revisions have been made by different people, including a version created by Jennifer Bayliss and Tim Garwood at Rochester Institute of Technology, Rochester, NY, for UT2004 that has a number of improvements. This version was further modified for Pogamut.

While *Unreal Tournament 2004* is fairly old by gaming standards, there is still an active modding community, and many online resources and guides are available. This allowed us to make our own changes to the game so that it would better suit the purposes of the competition. We will explain the changes that were made in Sections VI-D.

One of the "game types" supported in UT2004 is the "Death Match." In this game format, the aim is to have your character frag as many other characters as possible within the time limits of the game. We chose this game format as it lends itself well to the task. It is a popular choice among gamers. A Death Match with three participants (judge, confederate, and bot) makes sense, and a game can be completed in a relatively short time, which is important for practical reasons of time limitations in running the competition. In this format, the players are playing against each other rather than cooperatively. Choosing a cooperative game format would offer opportunities for interesting interactions, and will be considered for future competitions.

B. Competition Format

The competition was run in a number of rounds, similar to the Loebner Prize format. In each round, each judge was matched against a human player and a bot, and played a 10-min Death Match.

The human players, or confederates, were all moderately skilled players (self-reported). They were all male computer science students between 18 and 25 years of age. They were instructed to play the game as they normally would. Debriefing and an inspection of the game logs showed that they did so. As an incentive, they played for a small cash prize of $150, plus a trophy. Judges and competitors were given this information in advance. Therefore, competitors were able to aim to imitate such players, and judges were able to expect the confederates to be reasonably skilled and playing to win.

In the competition final, there were five judges, five confederates, and five bots. This allowed for a design in which five rounds were played, and each judge met each confederate and each bot exactly once over the five rounds. All the bots in the final had been pretested in a qualifying stage, ensuring that they

[2]http://artemis.ms.mff.cuni.cz/pogamut

TABLE I
SEQUENCE OF GAMES IN THE 2008 BOTPRIZE FINAL

	Judge 1		Judge 2		Judge 3		Judge 4		Judge 5	
Round	Bot	Human	Bot	Human	Bot	Human	Bot	Human	Bot	Human
1	1	1	5	2	3	3	2	4	4	5
2	2	2	3	1	4	5	5	3	1	4
3	3	3	4	5	5	4	1	2	2	1
4	4	4	1	3	2	1	3	5	5	2
5	5	5	2	4	1	2	4	1	3	3

worked correctly with the game software and would therefore have a reasonable *a priori* chance to win. (In fact, six teams qualified, but one was unable to make the trip to take part in the final.)

Table I shows the schedule that was used. After the competition, one of the competitors pointed out that it is possible to arrange the schedule so that each bot and human also meet exactly once during the competition. Such a schedule will be used in future.

The physical arrangement was to have one computer for each server, and one for each confederate, all set up in one room, and one computer for each judge, set up in another room, with no communication between the rooms other than by the organizers, or via game play. Spectators were able to come to the judges' room and watch the games in progress (see Fig. 3).

Prior to the final judging, the competing teams were given time to install and test their bots on the client machines. There was also a practice session set aside, during which the bots and confederates could learn to play the modified version of the game that was used in judging. No coding or other manual changes were allowed during the practice sessions, but the bots could use what they learned in practice to improve their play in the judging session. The same rules applied during the judging sessions: no manual adjustments were allowed.

The protocol for each judging round was as follows.

1) The servers were started.
2) A bot was started and connected to each server.
3) The confederates each connected to a server.
4) The judges were now allowed to join the game on their assigned server.
5) Each game was a 10-min Death Match.
6) At the end of the round, each judge was asked to rate the two opponents on a "humanness" scale, and to record their observations.
7) After a short break, the next round was begun.

This is the rating scale used by the judges.

1) This player is a not very humanlike bot.
2) This player is a fairly humanlike bot.
3) This player is a quite humanlike bot.
4) This play is a very humanlike bot.
5) This player is human.

Although judges were told that there would be a human and a bot opponent in each round, it was open to them to judge both opponents as human, or both as bots, if they wished. Note that this is different from the Turing Test, in which judges are forced to choose one or the other as human, and the other, by implication, as not human. A forced choice in effect compels judges to "toss a coin" if they really cannot pick the human.

In order to pass the test, a bot (or human) was required to achieve a rating of 4 (this player is human) from four of the five judges, i.e., 80%. If no bot passed the test, then the one with the highest mean humanness rating would be declared the competition winner.

The 80% may at first seem to be a stringent cutoff, but recall that in the Turing Test (and the Loebner Prize), the judges are forced to choose whether a player is human, whereas in the BotPrize, judges have the option to nominate only one player as human, or both players, or neither. If a human were allowed to enter the Loebner prize, judges would be forced to choose which of two apparently human players to nominate as the bot. Thus, a human could only expect 50% success in the Loebner Prize on average. However, if a human were allowed to enter the BotPrize contest, judges could nominate both players as human. Thus, a human could conceivably score 100% success in the BotPrize contest. Allowing for some error on the part of judges, 80% seems a reasonable threshold to use.

C. The Judges

The outcome of the competition depended, of course, on the abilities of the judges to tell which opponents were bots. Our judges were:

1) Penny Sweetser: a former AI researcher, now a senior game designer at 2K Australia (the competition sponsor);
2) Cam Atkinson: an expert level game player;
3) Robert J. Spencer (RJ): the chief operating officer of games company Interzone Entertainment (Perth, W.A., Australia);
4) John Wiese: a senior project manager for Special Operations Command Support Systems at defense contractor Thales (Perth, W.A., Australia);
5) David Fogel: a prominent AI researcher and President of the IEEE Computational Intelligence Society.

Of the five judges, all except David are avid and proficient interactive video game players. All except Cam have some knowledge of artificial intelligence methods, with at least Penny and David being expert in the use of artificial intelligence in games.

One weakness in the judging panel was the lack of any special expertise in human behavior. In future competitions, we will seek to include a cognitive scientist on the panel to address this weakness.

D. Competition Mods

Some mods were made to the standard game to facilitate the competition. There were two types of changes:

- changes to the game;
- changes to the semantics of GameBots messages (but not to the syntax).

Some of these changes were made to require bots to learn something about the game, as a human can do, rather than simply following predefined scripts. Although the definition of the test does not explicitly require learning, human players have to learn how best to play a new map and what strategies and tactics to use for modded versions of games. Thus, it was considered important to provide scope for bots to exhibit humanlike learning abilities (or the lack of them). In-game chatting was also disabled, as chatting would provide an easy way for judges to identify bots, and would change the nature of the competition, making it a chatbot competition. Other changes were to prevent bots from disrupting the competition. Changes were made to the following game features.

1) In-Game Visual Information: The 3-D representation of the game as seen by the players normally contains information that could be used to identify opponents based on their names. We modified the game to eliminate any possibility of judges using this to distinguish bots from human players, for example, by sharing information about different players between rounds.

The game was modified so that as each player, either human or bot, enters the game, their name is changed to a randomly generated one like "player265." This name is then displayed above the character during the game, and used in various game status displays that are available to players. The names therefore give judges no hint about the nature of an opponent, and information about a particular opponent cannot be transferred to subsequent rounds.

At the end of the game, the server wrote a list of players and their assigned names to file, so that the obfuscation could be untangled to analyze the results.

In addition, the appearance of characters in the games can be customized. This was standardized so that all characters had the same physical appearance, apart from their assigned random names.

2) The Map: The virtual world that the game simulates is described by what is called a "map." The game engine is capable of simulating different physical spaces defined by the map, which is typically loaded from file when the game is started. Tools called "map editors" can be used to create and edit maps. The map describes the 3-D physical layout of the world (for example, its dimensions, positions of walls, and other objects), other information about objects in the virtual world, as well as image and sound data used to create a real-time 3-D view (usually from the viewpoint of the character being controlled by the player) and to provide sound effects, character voices, and background music. Maps also usually contain information that is intended to be used by bots, to help them behave sensibly in the game. Maps in UT2004 typically include "hints" about locations and items. For example, a particular location might be labeled as being a good "sniping" spot, where a bot could hide and wait for opponents to pass by. This information is normally passed to bots via messages in the GameBots interface as was described in Section IV-A.

We modified GameBots to strip out these hints, to encourage competitors to have their bots learn this information for themselves, as a human player would have to do.

We also chose a map from the public domain that competitors would be unlikely to have seen before, to discourage them from hard-coding this information in their bots. GameBots usually provides bots with the name of the map, but we removed this information also. We chose a medium-sized map, so that judges could easily locate the bots within the game's time limit, yet players had room to run and hide from each other to recover health, replenish ammunition, lie in wait, and so on. When judges were killed, they respawned close to the main arena, so that there was little disruption to the judging. The map also had an unusual feature: some of the walls only appeared when a player attempted to pass through a particular volume of space.

An integral part of map is the network of navigation points, intended to be used by bots to make it simple for them to find their way around the virtual world. We considered removing this information from the GameBots interface, forcing bots to learn the physical layout for themselves, and make their navigation decision unaided. However, feedback from potential participants indicated that this would raise the entry barrier too high for this initial competition.

3) Weapons: In order to give the judges more time to observe the players, we reduced the damage done by weapons to 20% of normal.

UT2004 has a standard set of weapons available, each with its own capabilities. Modders can change the effects of these weapons, or define completely new weapons with new effects.

To further encourage competitors to have their bots learn this information, they were told that existing weapons may have their effects changed. In fact, due to the last minute discovery of a technical hitch, this modification was disabled, and the weapons had their usual effects.

GameBots normally provides bots with hints about the capabilities of weapons (for example, whether they are suited for use in a melee). These hints were removed. GameBots has a command that automatically selects the best weapon for a given situation. This command was changed so that it would select a random weapon. This was done to encourage competitors to have their bots make this choice for themselves.

4) Server Control: GameBots has a number of commands that can affect the game state (for example, switching to a new map, or pausing the game). These commands were disabled.

5) Combos: Competitors and confederates were asked not to make any use of combos, because GameBots did not allow bots to execute them. The latest version of GameBots does allow bots to execute combos, so these will be allowed in future.

E. The Competitors

Five bots took part in the final:

1) U Texas: entered by the team of Jacob Schrum, Igor Karpov, and Risto Miikkulainen of the Neural Networks Research Group, University of Texas at Austin;

2) AMIS: entered by Michal Stolba, Jakub Gemrot, and Juraj Simlovic, a group of students from Charles University, Prague, Czech Republic, who are also responsible for Pogamut;
3) Underdog: entered by Oren Nachman, a student at The University of Western Australia, Perth, W.A., Australia;
4) ICE: entered by Daichi Hirono, Yuna Dei and Ruck Thawonmas from the Department of Human and Computer Intelligence, Ritsumeikan University, Japan;
5) ISC: entered by Budhitama Subagdja, Ah Hwee Tan, and Di Wang from the Intelligent Systems Centre, Nanyang Technological University, Singapore.

Here we briefly describe each of these bots and the adaptations made to try to make them behave in a more humanlike way.

1) U Texas: This bot was based on an example bot, Hunter, which is included in Pogamut, but improved in several ways. The bot learned weapon preferences by keeping a table of expected damage per shot for each weapon, along with a table of each weapon's accuracy for three predefined ranges. It would choose the weapon where the product of the expected damage and accuracy for the current range were the largest. Message listeners were added to the bot to make it probabilistically dodge sideways in response to incoming projectiles. It also learned how best to strafe in combat by attempting to minimize health lost, and attempted to avoid running into walls by minimizing time spent colliding [26]. Note that, in general, bots do not stay exactly on navigation points, so it is possible to run into walls and get stuck in corners.

2) AMIS: The AMIS bot was a modified version of a previously created bot, which was designed to be "as good a death match player as possible" [27]. The changes made for the competition were to provide the ability to learn about weapons, and with the capacity to make mistakes.

To learn about weapons, the bot made use of a weapon information data structure that is part of its memory. The bot could be operated in a learning mode, during which it would play in such a way as to gather information about weapons available in the game, such as how much damage they do to opponents, depending on distance to target. The action selection mechanism of the bot uses a collection of drives (fight, pursue, etc.). Another drive to learn weapon information is added when learning mode is enabled.

Two kinds of "mistakes" were introduced to the bot's behavior. First, the bot team made a virtue of the bot's imperfect learning performance. The weapon information collected in learning mode is not sufficient to allow the bot to hit its target 100% of the time. For example, it was not possible to learn exactly how far to aim ahead of a moving opponent, which results in the bot missing shots (especially longer shots), as a human might be expected to do. The second deliberate weakness that was introduced was with movement. Rather than constantly strafing during combat, the bot would, with some small probability, stop moving, making it easier to hit. Similarly, with some probability, it would choose not to jump, even when jumping would have given the bot a better chance to avoid being hit.

The AMIS bot could thus be described as a very good player, deliberately made less proficient to make its actions seem more believably humanlike. More detailed description is available on the Pogamut web site.

3) Underdog: Underdog was designed to play as a reasonably good human, good enough to hold its own against bots and casual players but not good enough to beat a good player.

A number of "humanlike" qualities were introduced including the following.

1) Slight random delays were introduced into firing sequences, especially when firing from long distances. Although indistinguishable to human eye, these delays caused a slight loss of precision when firing at a moving target.
2) The bot would hunt down a player even if it lost eye contact with them.
3) The player would jump at close proximity, in order to avoid being hit.

The robot also attempted to learn about its opponents, getting slightly worse against players that it was consistently killing and attempting to get a little better against better players [28].

4) ICE: This bot has two states: exploring and fighting. In the exploring state, the bot's goal is to collect weapons and other items. In the fighting state, the bot's goal is to shoot the other players and to continuously learn the firing range of each weapon. The exploring state (also the initial state) changes to the fighting state if another character is seen. The fighting state changes to the exploring state if the target character disappears from the bot's sight. Some common tricks were also used to make the bot more believable, such as gradually decreasing the shooting precision for further targets. [29]

5) ISC: This bot employed FALCON and TD-FALCON, ART-based self-organizing neural network systems using reinforcement learning [31]. TD-FALCON was used to select from four states (exploring, collecting, escaping, and fighting) based on eight inputs (health level, whether suffering damage, whether an opponent is in view, whether ammunition level is adequate, plus the current state). The detailed behavior of the bot in each state was hard-coded, with a degree of randomness included. FALCON [32] was used for weapon selection (based on opponent distance). Real-time learning was used during play, so that the bot would be less predictable, and therefore, more humanlike.

In order to make the bot seem more human, a number of hopefully humanlike actions were embedded in the behavior states. Two examples were turning by a random amount to try to locate an opponent when hit, and choosing appropriately whether to try to jump over obstacles or to go around them.

VII. Results

As well as assigning a humanness rating to each player, judges were asked to offer comments. Both of these data sets are described below. An additional source of data is the videos made of all the games in the finals.[3] As well as the video, recordings were made of all the games in UT2004's proprietary "demo" format, which records all actions and events in the game (the videos were made by playing these back using

[3]Available at www.botprize.org

TABLE II
SUMMARY OF HUMAN AND BOT PERFORMANCE

Identity	Mean rating	#Judges convinced	Frags	Weighted mean
Human 3 (Byron)	4	5	74	4
Human 1 (Andrew)	3.8	4	65	3.96
Human 2 (Roderick)	3.8	4	91	3.7
Human 4 (Keith)	3	2	60	3.2
Human 5 (Seb)	2.8	2	82	3.28
Human mean	3.48	3.4	74.4	3.63
Bot 2 (AMIS)	2.4	2	58	2.28
Bot 4 (ICE)	2.2	1	52	1.59
Bot 5 (ISC)	2	2	93	1.83
Bot 1 (UTexas)	0.8	0	46	0.63
Bot 3 (Underdog)	0.4	0	65	0.46
Bot mean	1.56	1	62.8	1.36

UT2004's replay facility). These can be obtained from the author on request.

A. Ratings

Table II summarizes the performance of the confederates and bots over the five rounds. The rows of the table are sorted by mean humanness rating, with the most human at the top. (A complete table of results is given in Appendix I.)

None of the bots passed the test, that is, none fooled the required four out of five judges into giving them the fully human rating, but two of the five bots (AMIS and ISC) did manage to fool two judges. In addition, ICE fooled one judge.

It is reassuring to note that all the human confederates were judged to be more human than all the bots, giving us confidence in both the methodology of the test, and the ability of the judges to discriminate between human and bot behavior.

The fourth column of the table shows the number of frags obtained by each player over the five rounds. While by this measure, the human players did better than the bots on average, the highest scorer was the bot ISC, which was only the eighth most humanlike player, and only third most humanlike among the bots. The significance of this is discussed in the next section.

The final column shows the weighted average humanness rating for each player, where we have weighted the judge's ratings according to their accuracy. We calculated an accuracy figure for each judge as the mean rating the judge gave to human players minus the mean rating given to bots. The weighted means calculated using these weights lead to an almost identical ordering of the players, increasing our confidence that a meaningful ordering has been obtained.

TABLE III
SUMMARY OF JUDGING PERFORMANCE

Judge	Mean human rating	Mean bot rating	Accuracy	Humans correct	Bots correct
John	3.8	1	2.8	4	5
Cam	3.6	1	2.6	4	4
Penny	3.6	1.6	2	4	4
RJ	3.4	2.2	1.2	3	4
David	2.8	2	0.8	2	3

Table III tabulates the performance of the judges in correctly identifying bots and humans. It could be argued that the calculated accuracy rating is unfair, as the judges were not trying to maximize their accuracy score. However, the scores calculated do correlate well with the number of correct judgments made by the judges, so we feel justified in using it for the purpose of weighting judgments.

B. Judge's Comments

An edited transcript of the judges' comments can be found in Appendix II. In Table IV, we tabulate categories of characteristics reported by the judges in these comments (one "x" for each comment in that category). Of course, the allocation of comments to these categories is subjective. Some comments that seemed unclear or unrelated to other comments have been omitted from the table.

Categories between "Behaviors missing" and "Poor tactics" are characteristics that the judges used to identify players as bots. Those between "Skill errors/poor skills" and "Good tactics" were used to identify players as human. The entries in the shaded quadrants highlight mistaken judgments made by the judges.

VIII. DISCUSSION

We proposed and organized the BotPrize competition to stimulate discussion and research into how to make a bot appear like a human player. We believe that answers to this question will be important for artificial intelligence researchers, as well as for game developers. What answers did we get?

A. How Well Did the Test Work?

Before we proposed the competition, we discussed the idea with colleagues and acquaintances. While many were supportive in their comments, many others predicted that the competition would, for various reasons, be a failure.

A common prediction was that the judges would not be able to distinguish expert human players from bots, because expert humans play so well that their skills seem superhuman. We largely avoided this potential problem by choosing confederates who were all reasonably skilled players, but not elite players. We

made this choice known to competitors and judges, with the result that none of the bots was designed to imitate an elite player.

However, as far as the competition results go, they suggest that there is no correlation between skill levels and humanness, at least within this band of competent-but-not-expert skill. The most and least successful confederates (in terms of number of frags), were judged third and fourth most human, in the middle of the humans. Likewise, the most and least successful bots (again, in terms of number of frags), were judged third and fourth most human among the bots.

This contrasts with Laird and Duchi's result that skill level is a critical factor in appearing human. This may be because they tested a wider range of skill levels. In future competitions, we therefore intend to recruit confederates with a wider range of skill levels, from beginner to expert. Such change may make the test easier for bots to pass—bot designers could choose to imitate a beginner or an expert player, which may be easier than imitating a player with average skills.

After the competition, there was some interesting discussion about it.[4] Several contributors commented along the lines that it would be easy to imitate a 12-year-old player by moving and firing randomly and swearing whenever fragged (though that would not have worked as chatting was disabled). Perhaps we might see some bots with unusual behaviors in future competitions.

Perversely, another common prediction was that the judges would find the task too easy. These predictions were based on anecdotal evidence that bots can be identified either by their predictability, or by their inappropriate responses to unusual game conditions or player actions. The folk wisdom among game players is that it is easy to find and exploit such weaknesses to defeat bots.

We believe that the competition results indicate that the task is not easy. While the judges did well, their performance was far from perfect. The humanness scores were rather similar for the confederates who scored lowest on the humanness scale (we emphasize the very narrow sense of "humanness" here) and the bots that scored highest, and the winning bot was rated fully human by two of the five judges—the same number of fully human ratings as achieved by two of the confederates.

B. What Did We Learn?

Because of the practical requirements of the competition format, there is not enough data here for a rigorous statistical analysis of the results. Yet the data do suggest some interesting hypotheses for future testing.

First, we can analyze the comments to discover what characteristics were most reliable in distinguishing humans from bots.

From the judges' comments as categorized in Table III, we see that there were several characteristics whose presence always indicated that the player was a bot. These were:

- missing behaviors;
- stupid behaviors;
- low aggression.

[4]The discussion took place on the technology-related news website www.slashdot.org

The four examples of missing behaviors were: comments 1.1.2 (did not track me), 1.1.5 (did not dodge), 2.1.2 (limited range of reactions/strategies), and 3.5.1 (kept jumping vertically [no side jumps, like other humans]). One way to improve the humanness of these bots would be to add the missing behaviors, either by specifically coding them into the bot's repertoire (easy), or by giving the bots the ability to generate appropriate behaviors to suit the situation (hard).

Stupid behaviors noted by the judges were: comments 1.1.3 (stuck on wall), 1.4.1 (seemed to get confused by environment [i.e., did not respond naturally under fire]), 2.1.1 (jumped up and down on the spot to avoid incoming fire, did not move around), 3.1.2 (confused by different levels), and 4.4.1 (got stuck in corners). These specific problems could be corrected by specific code changes (easy), but one could never be sure that all the stupid behaviors had been eliminated—some unanticipated situation could cause another stupid behavior. Perhaps the only way to be sure would be to give the bot "real intelligence" (hard).

Judges successfully used the players' level of aggression to distinguish bots from humans: low aggression indicating a bot and high aggression indicating a human. Perhaps aggression is identified with humanness because it suggests determination and an emotional drive. It should be a relatively easy thing for bot programmers to make their bots more aggressive. A note of caution: it may be that aggressiveness is characteristic of humanness in first-person shooters, but not necessarily in other types of games.

Contrariwise, there were also characteristics whose presence (almost) always indicated that the player was human. These were:

- high aggression;
- adapting to the opponent;
- good tactics.

Three comments categorized as "adapting to the opponent" were: comments 6.4.1 (would stand back and watch what I did), 6.2.2 (was also slow to react to my movement—"anticipating" my moves), and 10.4.1 (learn where I like to go and beat me there). The judges here regard the player as forming some kind of opponent model, a level of awareness that they associate with being human. This may be one area where bot programmers could use some quite simple opponent modeling methods to give the appearance of awareness.

On the subject of tactics, good tactics were taken as an indicator of humanness, and bad tactics as an indicator of "botness" (although one human was thought to show poor tactics by one judge). This contrasts with the skills categories, where good skills were often judged to be too good, indicating botness, and skill errors were often judged to indicate humanness (although these judgments were not always right; see below).

Perhaps this indicates that the judges expect bots to perform poorly at tasks requiring intelligence (tactics) and to excel at tasks requiring precise timing and calculation (skills). Examples of good tactics were: comments 7.1.2 (moved around to dodge/avoid fire well), 7.1.4 (combos with movement/weapons), 8.1.2 (used different weapons well), 8.1.3 (strafed around), 8.1.4 (used different heights properly), 8.5.2 (when I was engaged with [the other player] came and attacked

TABLE IV
PLAYER CHARACTERISTICS AS REPORTED IN THE JUDGES' COMMENTS

Category	Bots					Humans				
	1	2	3	4	5	1	2	3	4	5
Behaviors missing	xx	x	x							
Stupid behavior	xx	x	x	x						
Nonhuman behavior		x	xxxx	x	x		x		x	
Consistent / predictable	xx	x						x		xx
High skills	x		x	xx	x			x		xx
Low aggression			x							
Poor tactics					xx		x			
Skill errors/ poor skills	x	x			x		x	xx	x	
High aggression						xx	xx	xx	x	
Varied / unpredictable						x				
Adapting to opponent						xx				x
Good tactics	x					xx	xxxx	xxx	x	

Fig. 3. Judge's room. The judges are at the front of the room. One of the games is being displayed on the large screen for the benefit of spectators.

me), and 9.4.1 (chased when they outgunned you), while examples of bad tactics were: comments 5.4.1 (bot used environment poorly), 5.4.2 (did not target most dangerous player), and 7.4.2 (also wrong use of weapons at times which makes me think it was not human).

We can also examine the comments to discover which of the characteristics used by the judges were the least reliable, i.e., those that led the judges to make mistakes. The most misleading categories were:

- nonhuman behavior;
- consistent/predictable;
- skill level.

The following behaviors were correctly judged nonhuman: comments 2.1.1 (jumped up and down on the spot to avoid incoming fire, did not move around), 3.1.1 (jumped up and down too much, less strafing), 3.2.2 (often jumped on the spot), 3.4.2 (too much jumping on the spot with very good aim), 3.5.1 (kept jumping vertically [no side jumps, like other humans]), 4.2.1 (always ran in straight lines), and 5.2.1 (had 100% accuracy with instant hit weapons and usually stood still jumping), while the following were incorrectly judges nonhuman: 7.4.1 ([human 2] was very good but ran in perfect circles around me) and 9.5.1 ([human 4] acted quite robotically in engagements that I saw with [the other player]). A lot of these comments were made about Underdog, nicknamed "the hoppy bot" by onlookers. This bot jumped up and down on the spot too much, due to a bug inadvertently introduced in last-minute tuning, so we should perhaps discount this data to a degree.

Judges attempted to use predictability to identify bots, often unsuccessfully. Anecdotally, predictability is often touted as being a characteristic of bots, and an earlier study [10] found that predictability is strongly correlated with the perception that a player is a bot. It is interesting to note also that most of the teams attempted to make their bots unpredictable in some way.

Skill level was another area of confusion. Judges were sometimes unwilling to accept that a human opponent was highly skilled, and bot makers were often able to trick the judges by the simple means of incorporating deliberate skill errors in their bots.

One of the judges (John Weise) wrote of his experiences on his web site [33]. The reader will find that he reinforces some of our findings, as well as offers his own analysis. In one example, he relates how he discovered a tactic peculiar to the map used in

the final, and used it to evade bots. Humans, however, were able to figure out what he was doing, and work out a countertactic. This example illustrates the fact that the judges developed tactics designed to uncover the bots, similar to the judges in the Loebner Prize.

IX. Future Work

Over time, we expect the bots to become more difficult to distinguish from the human confederates (just as the entries in the Loebner Prize seem to have improved), and we expect to be able to include more challenging problems for bot designers to deal with.

For example, the bots' perceptions of the game are currently limited to a degree by the information that the GameBots interface provides. Imagine that a modder creates a mod in which one of the weapons can be used to temporarily turn a section of the floor into an impassable lava pit. A human player would quickly be able to recognize, using vision, that a lava pit has appeared, and would be able to infer likely effects of entering the pit, and would then be able to plan how to make use of this new feature to gain an advantage or avoid damage. But the GameBots interface does not provide a visual representation of the game to bots, and has no message that indicates the presence of a lava pit (the concept is not even in the GameBots vocabulary).

A distant future goal is to replace the GameBots interface with the same visual, acoustic, and tactile input that is provided to human players. Admittedly, the capability of bots to use these inputs is a long way away, but some researchers have made initial steps in this direction [34].

A less distant possibility is to provide some limited chatting capability, so that bots could taunt or annoy opponents, adding an emotional dimension to the task. (Unrestricted chatting is unlikely to be allowed any time soon; someone would have to win the Loebner prize first.)

Another possibility is to remove the navigation map data entirely, requiring bots to learn the layout of the virtual game world through trial and error. Techniques developed for map learning in robotic applications might be adapted to do this [35].

One of our anonymous reviewers pointed out the potential importance of social context and narrative in influencing players' perception of other players' humanness (as argued in, for example, [36]). The current format of the test restricts the opportunity for such factors to come into play. Future competitions will look to enrich the social context where possible. Introducing restricted chatting, longer rounds, and team play are future possibilities.

We believe that there is much more that can be learned from the Turing Test for Bots and we hope to run the BotPrize Contest annually. At the time of writing, the 2009 competition is confirmed and will take place at the 2009 IEEE Symposium on Computational Intelligence and Games in Milano, Italy, in September.[5] Fifteen research teams have announced their intention to enter. It will be interesting to see how the bots have progressed.

[5]Details are now available at www.botprize.org

X. Conclusion

The overall result of the 2008 competition could be summarized as follows:

Computers cannot play like humans—yet.

The test allowed the judges to discriminate between human and bot players, with all humans being judged more human than all bots. Three out of five humans passed the test, and none of the bots did. The bots did have some success, however, with two bots deceiving two judges, and another bot deceiving one judge.

The bot designers had most success in confusing the judges by:

- using (pseudo)randomness to simulate unpredictability;
- incorporating skill errors in the bots' play.

Based on the judges' comments, some opportunities to make bots more convincing as imitators are:

- increasing the range of humanlike behaviors that bots exhibit;
- eliminating obviously stupid behaviors;
- increasing bots' apparent aggression levels;
- appearing to adapt play to suit the opponent;
- exhibiting sound tactical play.

All of these, except for increased aggression, relate to the bots' apparent intelligence—a link back to the original Turing test. For example, repeating an "obviously stupid" behavior is evidence of a lack of true intelligence (whatever that might be): a human would be sufficiently aware as to be able to recognize such behavior as stupid, and change it. Both the recognition task and the task of synthesizing an alternative behavior might be expected to require a high degree of intelligence. Even aggression can be taken as an indication of a mental process taking place—the pursuit of a goal (killing the opponent) over an extended period.

Some of these improvements may be easily "faked," while others may require more complex solutions, equipping bots with a deeper "understanding" of their environment and their task, and with the ability to learn and adapt—traditional goals of the artificial intelligence researcher. From an artificial intelligence perspective, the fact that many of the important behaviors are cognitive ones is encouraging.

As a challenge problem, we are optimistic that the BotPrize Contest will be an effective and fruitful one. The evidence from the 2008 Contest is that the task is just a little beyond the current state of the art: neither so easy that we need not stretch ourselves to succeed, nor so hard that success cannot be imagined. Competitors have begun to publish papers describing their methods. We look forward with some eagerness to the next competition.

Appendix I
Table of Humanness Ratings

Table V tabulates the humanness ratings round by round. Each row gives the results for a particular round, for the game involving a particular bot, and the rows for each bot are grouped together in round order. The column labeled "Bot Score" (resp.,

TABLE V
TABULATED RESULTS FROM THE 2008 BOTPRIZE CONTEST, ROUND BY ROUND

Round	Judge	Human	Bot Rating	Bot Score	Bot Deaths	Human Rating	Human Score	Human Deaths
Bot 1 (U Texas)								
1	1	1	0	2	6	4	8	1
2	5	4	1	3	6	2	7	6
3	4	2	0	4	7	3	15	3
4	2	3	1	4	5	4	3	6
5	3	2	2	1	8	4	16	3
Bot 2 (AMIS)								
1	4	4	2	1	10	4	13	0
2	1	2	1	0	13	4	15	0
3	5	1	4	1	14	3	21	0
4	3	1	1	0	12	4	15	0
5	2	4	4	1	6	2	5	4
Bot 3 (Underdog)								
1	3	3	1	6	2	4	5	5
2	2	1	0	3	10	4	7	2
3	1	3	0	4	6	4	8	6
4	4	5	1	8	10	4	15	4
5	5	3	0	7	9	4	20	5
Bot 4 (ICE)								
1	5	5	4	3	10	1	16	0
2	3	5	3	2	8	2	12	1
3	2	5	0	2	9	4	10	2
4	1	4	3	2	6	4	10	0
5	4	1	1	3	7	4	11	0
Bot 5 (ISC)								
1	2	2	0	6	11	4	9	4
2	4	3	1	17	7	4	7	9
3	3	4	4	5	9	3	11	4
4	5	2	1	2	17	4	25	1
5	1	5	4	4	15	2	20	2

"Human Score") is the number of frags obtained by the bot (resp., human) in that round minus the number of times the bot (resp., human) was fragged. The shaded cells are those where the judge has made an incorrect identification.

APPENDIX II
TRANSCRIPTS OF JUDGES' COMMENTS

Here are transcriptions of the judge's comments. The raw comments referred to the players by their (randomly generated) names. We have replaced these names with the players' true identities (which were unknown to the judges). We have also rearranged the comments into a list of comments on each player.

1) Bot 1: U Texas (mean humanness rating 0.8)
   - 1.1) Judge 1: Penny (accuracy $= 2$) – rating $= 0$
     - 1.1.1) Shooting too even.
     - 1.1.2) Did not track me.
     - 1.1.3) Stuck on wall.
     - 1.1.4) Behavior too consistent.
     - 1.1.5) Did not dodge.
   - 1.2) Judge 2: Cam (accuracy $= 2.6$) – rating $= 1$
     - 1.2.1) Moved well but was 100% accurate with instant hit weapons.
   - 1.3) Judge 3: RJ (accuracy $= 1.4$) – rating $= 2$
     - 1.3.1) RJ made no specific comment here, but noted that "confederates that were too good/aggressive made it hard to determine the value of the bot," i.e., he knew which was the bot, but found it hard to decide on a rating.
   - 1.4) Judge 4: John (accuracy $= 2.8$) – rating $= 0$
     - 1.4.1) Seemed to get confused by environment (i.e., did not respond naturally under fire).
   - 1.5) Judge 5: David (accuracy $= 0.4$) – rating $= 1$
     - 1.5.1) I actually beat (Bot 1) in this round. Since I am so bad, I have to presume that (Bot 1) is not human. Any human should beat me.
     - 1.5.2) Neither played very well, but I would say (the other player) was more human. But it would really be impossible for me to say either *is* human.
2) Bot 2: AMIS (mean humanness rating 2.4)
   - 2.1) Judge 1: Penny (accuracy $= 2$) – rating $= 1$
     - 2.1.1) Jumped up and down on the spot to avoid incoming fire, did not move around.
     - 2.1.2) Limited range of reactions/strategies.
   - 2.2) Judge 2: Cam (accuracy $= 2.6$) – rating $= 4$
     - 2.2.1) Fired erratically and led the player around instead of fighting.
     - 2.2.2) Note: Cam thought the other player was a bot because it appeared to keep tracking him even though he was invisible (Cam had used a combo to make himself invisible). This is probably because the competition mod would show the name of the player, in the correct location, even if the player was invisible. Cam would not have known this.
   - 2.3) Judge 3: RJ (accuracy $= 1.4$) – rating $= 1$
     - 2.3.1) No specific comment.
   - 2.4) Judge 4: John (accuracy $= 2.8$) – rating $= 2$
     - 2.4.1) Seemed a little slow to react.
   - 2.5) Judge 5: David (accuracy $= 0.4$) – rating $= 4$
     - 2.5.1) ... (Bot 2) lobbed a flak at me from far away early, and did nothing to convince me after that, that (Bot 2) was not human.
     - 2.5.2) (Bot 2) if human, is as bad as I am. I doubt that is possible, but I have to go with my intuition.
3) Bot 3: Underdog (mean humanness rating 0.4)
   - 3.1) Judge 1: Penny (accuracy $= 2$) – rating $= 0$
     - 3.1.1) Jumped up and down too much, less strafing.
     - 3.1.2) Confused by different levels.
     - 3.1.3) Not aggressive enough.
   - 3.2) Judge 2: Cam (accuracy $= 2.6$) – rating $= 0$
     - 3.2.1) Could still see and hit player with invisibility active.
     - 3.2.2) Often jumped on the spot.
   - 3.3) Judge 3: RJ (accuracy $= 1.4$) – rating $= 1$
     - 3.3.1) No specific comment.
   - 3.4) Judge 4: John (accuracy $= 2.8$) – rating $= 1$
     - 3.4.1) Very early looked like a bot.
     - 3.4.2) Too much jumping on the spot with very good aim.
   - 3.5) Judge 5: David (accuracy $= 0.4$) – rating $= 0$
     - 3.5.1) Kept jumping vertically (no side jumps, like other humans).
     - 3.5.2) If (Bot 3) is a human, the human is playing like a bot.
4) Bot 4: ICE (mean humanness rating 2.2)
   - 4.1) Judge 1: Penny (accuracy $= 2$) – rating $= 3$
     - 4.1.1) Seemed a bit too perfect, range at which it could spot me and snipe from, reaction time, etc.
     - 4.1.2) Penny made a general comment: "Hardest to pick, were very similar. Both seemed like good players."
   - 4.2) Judge 2: Cam (accuracy $= 2.6$) – rating $= 0$
     - 4.2.1) Always ran in straight lines.
   - 4.3) Judge 3: RJ (accuracy $= 1.4$) – rating $= 3$
     - 4.3.1) No specific comment.
   - 4.4) Judge 4: John (accuracy $= 2.8$) – rating $= 1$
     - 4.4.1) Got stuck in corners.
     - 4.4.2) Moved poorly but very good shot.
   - 4.5) Judge 5: David (accuracy $= 0.4$) – rating $= 4$
     - 4.5.1) David commented that the other player seemed more automated.
5) Bot 5: ISC (mean humanness rating 2)
   - 5.1) Judge 1: Penny (accuracy $= 2$) – rating $= 4$
     - 5.1.1) Seemed like an OK human player.
   - 5.2) Judge 2: Cam (accuracy $= 2.6$) – rating $= 0$
     - 5.2.1) Had 100% accuracy with instant hit weapons and usually stood still jumping.
   - 5.3) Judge 3: RJ (accuracy $= 1.4$) – rating $= 4$
     - 5.3.1) No comment.
   - 5.4) Judge 4: John (accuracy $= 2.8$) – rating $= 1$
     - 5.4.1) Bot used environment poorly.
     - 5.4.2) Did not target most dangerous player.
   - 5.5) Judge 5: David (accuracy $= 0.4$) – rating $= 1$
     - 5.5.1) (Bot 5) was worse than me. If this is a human, I think he/she and I both need more practice.
6) Human 1: Andrew (mean humanness rating 3.8)

6.1) Judge 1: Penny (accuracy = 2) – rating = 4
- 6.1.1) More varied and erratic.
- 6.1.2) Aggressively followed me.
- 6.1.3) Dodged shots.

6.2) Judge 2: Cam (accuracy = 2.6) – rating = 4
- 6.2.1) Effectively countered attacks using agility and weapons like the shield gun alt-fire.
- 6.2.2) Was also slow to react to my movement—"anticipating" my moves.

6.3) Judge 3: RJ (accuracy = 1.4) – rating = 4
- 6.3.1) No comment.

6.4) Judge 4: John (accuracy = 2.8) – rating = 4
- 6.4.1) Would stand back and watch what I did.

6.5) Judge 5: David (accuracy = 0.4) – rating = 3
- 6.5.1) (Human 1) was very effective. Could have been a good human.
- 6.5.2) Was ruthless like a good human.

7) Human 2: Roderick (mean humanness rating 3.8)

7.1) Judge 1: Penny (accuracy = 2) – rating = 4
- 7.1.1) Did not see me when I was near them.
- 7.1.2) Moved around to dodge/avoid fire well.
- 7.1.3) Pushed me aggressively.
- 7.1.4) Combos with movement/weapons.

7.2) Judge 2: Cam (accuracy = 2.6) – rating = 4
- 7.2.1) Showed no concern for self preservation and would often follow the player in a straight line.

7.3) Judge 3: RJ (accuracy = 1.4) – rating = 4
- 7.3.1) No specific comment here, but noted that "confederates that were too good/aggressive made it hard to determine the value of the bot," i.e., he knew which was the bot, but found it hard to decide on a rating.

7.4) Judge 4: John (accuracy = 2.8) – rating = 3
- 7.4.1) (Human 2) was very good but ran in perfect circles around me.
- 7.4.2) Also wrong use of weapons at times which makes me think it was not human.

7.5) Judge 5: David (accuracy = 0.4) – rating = 4
- 7.5.1) (Human 2) played very well and seemingly with tactics.
- 7.5.2) If (Human 2) is a bot, it can do a good job slaughtering two bad humans.
- 7.5.3) I have to go with (Human 2) as the human.

8) Human 3: Byron (mean humanness rating 4)

8.1) Judge 1: Penny (accuracy = 2) – rating = 4
- 8.1.1) Followed aggressively.
- 8.1.2) Used different weapons well.
- 8.1.3) Strafed around.
- 8.1.4) Used different heights properly.

8.2) Judge 2: Cam (accuracy = 2.6) – rating = 4
- 8.2.1) Found it difficult to track other players.
- 8.2.2) Attempted acrobatic moves such as wall dodging, but sometimes failed to do them properly.

8.3) Judge 3: RJ (accuracy = 1.4) – rating = 4
- 8.3.1) No specific comment.

8.4) Judge 4: John (accuracy = 2.8) – rating = 4
- 8.4.1) No comment.

8.5) Judge 5: David (accuracy = 0.4) – rating = 4
- 8.5.1) (Human 3) played very humanlike.
- 8.5.2) When I was engaged with (the other player) came and attacked me.
- 8.5.3) Seemed consistent.
- 8.5.4) Won easily.

9) Human 4: Keith (mean humanness rating 3)

9.1) Judge 1: Penny (accuracy = 2) – rating = 4
- 9.1.1) Seemed to make human errors and oversights.
- 9.1.2) Less perfect.

9.2) Judge 2: Cam (accuracy = 2.6) – rating = 3
- 9.2.1) Did not lose track of an invisible player.

9.3) Judge 3: RJ (accuracy = 1.4) – rating = 3
- 9.3.1) No comment.

9.4) Judge 4: John (accuracy = 2.8) – rating = 4
- 9.4.1) Chased when they outgunned you.

9.5) Judge 5: David (accuracy = 0.4) – rating = 2
- 9.5.1) (Human 4) acted quite robotically in engagements that I saw with (the other player).
- 9.5.2) Neither played very well.
- 9.5.3) It would really be impossible for me to say either is human.

10) Human 5: Seb (mean humanness rating 2.8)

10.1) Judge 1: Penny (accuracy = 2) – rating = 2
- 10.1.1) Seemed too fast and perfect.
- 10.1.2) Fire fast and consistently and tracked too well.

10.2) Judge 2: Cam (accuracy = 2.6) – rating = 4
- 10.2.1) Was highly accurate and agile, but seemed to react to items as they spawned (with the assumption it was not a player reacting to a sound).
- 10.2.2) Cam changed his rating for Seb from 3 to 4.

10.3) Judge 3: RJ (accuracy = 1.4) – rating = 2
- 10.3.1) No specific comment.

10.4) Judge 4: John (accuracy = 2.8) – rating = 4
- 10.4.1) Learn where I like to go and beat me there.

10.5) Judge 5: David (accuracy = 0.4) – rating = 1
- 10.5.1) (Human 5) appeared to me to be more "heuristic" based on its engagements.
- 10.5.2) This seemed automated to me.
- 10.5.3) It could be that it was just a good human taking advantage of me.

ACKNOWLEDGMENT

The author would like to thank the confederates and judges for volunteering their time and expertise. He would also like to thank the competitors and other colleagues for many interesting and helpful discussions. Finally, he would like to thank 2K Australia for its sponsorship of the competition.

REFERENCES

[1] A. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[2] J. Hutcheons, "Pseudo-intelligence and entertainment," in *Proc. IEEE Symp. Comput. Intell. Games*, Perth, W.A., Australia, May 3, 2009, CD-ROM.
[3] H. Loebner, Rules for the 2009 Loebner Prize, 2009 [Online]. Available: http://loebner.net/Prizef/2009_Contest/LP_2009.html
[4] M. Bida, Gamebots API List, 2009 [Online]. Available: http://artemis.ms.mff.cuni.cz/pogamut/tiki-index.php?page=Gamebots+API+list
[5] R. C. Arkin, *Behavior-Based Robotic*. Cambridge, MA: MIT Press, 1998.
[6] A. Khoo, G. Dunham, N. Trienens, and S. Sood, "Efficient, realistic NPC control systems using behavior-based techniques," in *Proc. AAAI Spring Symp. Ser., Artif. Intell. Interactive Entertain.*, Menlo Park, CA.
[7] V. K. Tatai and R. R. Gudwin, "Using a semiotics-inspired tool for the control of intelligent opponents in computer games," in *Proc. IEEE Int. Conf. Integr. Knowl. Intensive Multi-Agent Syst.*, Cambridge, MA, Sep. 30–Oct. 4 2003, pp. 647–652.
[8] S. Hladky and V. Bulitko, "An evaluation of models for predicting opponent locations in first-person shooter video games," in *Proc. IEEE Symp. Comput. Intell. Games*, Perth, W.A., Australia, Dec. 2008, pp. 39–46.
[9] J. E. Laird and J. C. Duchi, "Creating human-like synthetic characters with multiple skill levels: A case study using the Soar Quakebot," in *Proc. AAAI Fall Symp. Simulating Human Agents*, 2000, pp. 75–79, FS-0A-03.
[10] J. E. Laird, A. Newell, and P. S. Rosenbloom, "Soar: An architecture for general intelligence," *Artif. Intell.*, vol. 33, no. 3, pp. 1–64, 1987.
[11] B. Gorman, C. Thurau, C. Bauckhage, and M. Humphrys, "Believability testing and Bayesian imitation in interactive computer games," in *Proc. Int. Conf. Simul. Adapt. Behav.*, 2006, vol. 4095, pp. 655–666.
[12] B. Soni and P. Hingston, "Bots trained to play like a human are more fun," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Hong Kong, Jun. 2008, pp. 363–369.
[13] P. A. Saygin, I. Cicekli, and V. Akman, "Turing test: 50 years later," *Minds Mach.*, vol. 10, no. 4, pp. 463–518, 2000.
[14] S. M. Shieber, "Lessons from a restricted Turing test," *Commun. ACM*, vol. 37, pp. 70–78, 1994.
[15] H. G. Loebner, "In response," *Commun. ACM*, vol. 37, pp. 79–82, 1994.
[16] R. L. Purtill, "Beating the imitation game," *Mind*, vol. 80, pp. 290–294, 1971.
[17] K. Gunderson, "The imitation game," *Mind*, vol. 73, pp. 234–245, 1964.
[18] D. Dennett, "Can machines think?," in *How we Know*, M. Shafto, Ed. San Francisco, CA: Harper & Row, 1985, pp. 121–145.
[19] P. H. Millar, "On the point of the imitation game," *Mind*, vol. 82, pp. 595–597, 1973.
[20] N. Block, "Psychologism and behaviorism," *Philosoph. Rev.*, vol. 90, pp. 5–43, 1981.
[21] N. Block, "The mind as the software of the brain," in *An Invitation to Cognitive Science*, D. Osherson, L. Gleitman, S. Kosslyn, E. Smith, and S. Sternberg, Eds. Cambridge, MA: MIT Press, 1995.
[22] J. R. Searle, "Minds, brains and programs," *Behav. Brain Sci.*, vol. 3, pp. 417–424, 1980.
[23] D. Michie, "Turing's test and conscious thought," in *Machines and Thought: The Legacy of Alan Turing*, P. Millican and A. Clark, Eds. Oxford, U.K.: Oxford Univ. Press, 1996, pp. 27–51.
[24] R. French, "Subcognition and the limits of the Turing test," *Mind*, vol. 99, no. 393, pp. 53–65, 1990.
[25] R. Kadlec, J. Gemrot, O. Burkert, M. Bída, J. Havlíček, and C. Brom, "POGAMUT 2—A platform for fast development of virtual agents' behaviour," in *Proc. CGAMES*, La Rochelle, France, 2007, pp. 49–53.
[26] R. Miikkulainen, *Personal Communication.* May 2009.
[27] M. Stolba, Botprize 2008 Winning Bot, 2009 [Online]. Available: http://artemis.ms.mff.cuni.cz/pogamut/tiki index.php?page=Botprize%202008%20winning%20bot
[28] O. Nachman, *Personal Communication.* May 2009.
[29] R. Thawonmas, *Personal Communication.* Dec. 2009.
[30] D. Wang, B. Subagdja, A.-H. Tan, and G.-W. Ng, "Creating human-like autonomous players in real-time first person shooter computer games," in *Proc. 21st Annu. Conf. Innovat. Appl. Artif. Intell.*, Pasadena, CA, Jul. 14–16, 2009, pp. 173–178.
[31] A.-H. Tan, N. Lu, and D. Xiao, "Integrating temporal difference methods and self-organizing neural networks for reinforcement learning with delayed evaluative feedback," *IEEE Trans. Neural Netw.*, vol. 19, no. 2, pp. 230–244, Feb. 2008.
[32] A.-H. Tan, "FALCON: A fusion architecture for learning, Cognition, and navigation," in *Proc. Int. Joint Conf. Neural Netw.*, 2004, pp. 3297–3302.
[33] J. Weise, Comments on the 2008 BotPrize Contest, 2009 [Online]. Available: http://www.wiesej.com/2008/botprize.htm
[34] M. Parker and B. D. Bryant, "Visual control in Quake II with a cyclic controller," in *Proc. IEEE Symp. Comput. Intell. Games*, Perth, W.A., Australia, Dec. 2008, pp. 151–158.
[35] C. Stachniss, *Robotic Mapping and Exploration*. New York: Springer-Verlag, 2009, vol. 55, Springer Tracts in Advanced Robotics, p. XVIII.
[36] M. Mateas and A. Stern, "A behavior language for story-based believable agents," *IEEE Intell. Syst.*, vol. 17, no. 4, pp. 39–47, Jul. 2002.

Philip Hingston (M'00–SM'06) received the B.Sc. degree in pure mathematics from the University of Western Australia, Perth, W.A., Australia, in 1978 and the Ph.D. degree in mathematical logic from Monash University, Melbourne, Vic., Australia, in 1984.

Currently, he is an Associate Professor in the School of Computer and Information Science, Edith Cowan University, Perth, W.A., Australia. He was previously a Senior Research Scientist with Rio Tinto Plc. He recently coedited *Design By Evolution* (New York: Springer-Verlag, 2008). His research interests are in artificial intelligence, especially evolutionary computing, and its applications, particularly in industrial optimization and computer games

Dr. Hingston is a member of the ACM. He serves on a number of IEEE Computational Intelligence Society committees, chairing the task force on coevolution, and is an associate editor of the IEEE TRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI IN GAMES.