Int J Artif Intell Educ (2017) 27:353–384
DOI 10.1007/s40593-017-0144-1

ARTICLE

AI in Informal Science Education: Bringing Turing Back to Life to Perform the Turing Test

Avelino J. Gonzalez[1] · James R. Hollister[1] · Ronald F. DeMara[2] · Jason Leigh[3] · Brandan Lanman[4] · Sang-Yoon Lee[3] · Shane Parker[1] · Christopher Walls[1] · Jeanne Parker[1] · Josiah Wong[1] · Clayton Barham[1] · Bryan Wilder[1]

Published online: 16 March 2017

Abstract This paper describes an interactive museum exhibit featuring an avatar of Alan Turing that informs museum visitors about artificial intelligence and Turing's seminal Turing Test for machine intelligence. The objective of the exhibit is to engage and motivate visiting children in the hope of sparking an interest in them about computer science and artificial intelligence, and cause them to consider pursuing future studies and/or careers in these fields. The exhibit interacts with the visitors, allowing them to participate in a simplified version of Turing's test that is brief and informal to suit the limitations of a five-minute exhibit. In this exhibit, the visitor (targeted towards middle school age children) invokes an avatar of his/her own choice, and acts to endow it with human-like qualities (voice, brain, eyesight and hearing). Then, the visitor engages the avatar in a (brief) question-and-answer session to determine whether the visitor thinks that he/she is interacting with a real human on a video conference or with an avatar. We consider this interaction to be an extension of the original Turing Test because, unlike Turing's original that used text via a teletype, this version features a graphical embodiment of an agent with which one can converse in spoken natural language. This extension serves to make passing the Turing Test more difficult, as now the avatar must not only communicate like a human, but also look, sound and act the part. It also makes the exhibit visual, dynamic and interesting to the visitors. Evaluations were performed with museum visitors, both in backrooms with prototypes as well as on the museum floor with the final version. The

✉ Avelino J. Gonzalez
Avelino.gonzalez@ucf.edu

[1] Computer Science Department, University of Central Florida, Orlando, FL, USA

[2] Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL, USA

[3] Electronic Visualization Laboratory, University of Illinois at Chicago, Chicago, IL, USA

[4] Orlando Science Center, Orlando, FL, USA

formative and summative evaluations performed indicated overall success in engaging the museum visitors and increasing their interest in computer science. More specifically, the formative testing, mostly done in quiet back rooms with selected test subjects, indicated that on the important questions about enjoyment of exhibit and increased interest in computer science by the test subjects, their self-reported Likert scale responses (1 being negative and 5 being positive) increased from 3.16 in the first evaluation to 4.38 in the third one for increased interest in CS. Likewise for the question about exhibit enjoyment (from 3.92 to 4.56). The summative evaluation, done through unobtrusive observation of exhibit use on museum floor, indicated that almost 74% of the parties that initiated the exhibit were either highly or moderately engaged by the exhibit. However, there was one major negative finding, namely the overly long duration of the exhibit, which may have caused premature abandonment of the exhibit in several cases during the summative evaluation. These tests and their results are presented and discussed in detail in this paper. The exhibit has been on permanent display at the Orlando (FL) Science Center since June 2014 and has received a strongly positive response from visitors since that time.

Keywords Artificial virtual humans · Avatars · Embodied conversational agents · Informal science education · Science museum exhibit · Turing test

Introduction

Informal science education (ISE) has been a highly effective vehicle for attracting children to science and technology from an early age. Science museums offer highly-visual and hands-on exhibits that can bring to life many concepts in science and technology. Therefore, exhibits in science museums are often directed towards children, and offer fun and engaging ways to participate, motivate, and spark their interest in science, technology, engineering and mathematics, the so-called STEM subjects.

In their 2009 book on Informal Science Education, Bell et al. [2009] proposed "*a 'strands of science learning' framework that articulates science-specific capabilities supported by informal environments*" [Bell et al. 2009, p. 3]. They suggest six such strands. Discussion of these strands is beyond the scope of this narrative, but the first strand, which they identify as being of "... *particular relevance* ..." to ISE, states: "*Strand 1: Experience excitement, interest, and motivation to learn about phenomena in the natural and physical world*." [Bell et al. 2009, p. 4]. This is the primary motivation for our exhibit described here.

Furthermore, it is no secret that the US currently faces a shortage of professionals in the STEM disciplines. In their 2005 report "*Rising Above the Gathering Storm…*" [NRC 2005], the US National Research Council speaks of an alarming decline in the STEM workforce in the US, one that still rings true today. While we have lived with a shortage of engineers for many years now, the growing pervasiveness of technology in our everyday lives has increased the need for science and technology professionals. Yet, there is currently a low level of interest in STEM careers on the part of children who will make up the next generation work force. A study reported in *Science Educator* [Sorge 2007] found that a positive attitude toward science peaks in children at about age nine. Then, over the next five years, the student's attitude about science experiences a sharp and permanent decline of nearly 40%. Furthermore, Falk stated in a webinar [Falk 2011] that if by the 8th

grade, a child has not decided to pursue a career in science (STEM), the probability is high that he/she never will. On the other hand, if she/he does choose science by the 8th grade, chances are quite good that she/he will obtain a degree in a STEM field.

To engage and motivate a young audience, exhibits must present interesting topics in an exciting manner. This typically requires making the experience participatory. We believe that artificial intelligence can be an exceptionally interesting topic for children. This is especially true with the emergence in recent years of intelligent virtual players in video games (called "AIs") and smart conversational assistants such as Apple's Siri® and Microsoft's Cortana®. We therefore assert here that wrapping an exhibit around *intelligent virtual humans* (also called *Embodied Conversational Agents* or *avatars*) that can interact with a young museum visitor in spoken natural language (with a touch screen and/or text alternative) injects a strong element of excitement into the exhibit.

In this paper we present and describe a museum exhibit that interacts with a visitor to "create" an avatar with knowledge about a specific topic. This avatar then engages in a short Q&A session with the visitor that serves as a test for machine intelligence. This exhibit was originally (and partially) described in Hollister et al. [2013]. At the core of this exhibit is the integration of computer graphics with artificial intelligence. The centerpiece of the exhibit is an avatar of Dr. Alan Turing, the famed British mathematician, WWII code breaker and pioneer of modern computing [DiSalvo 2012]. Turing's avatar communicates with exhibit visitors in spoken natural language (or alternatively, typed text or a touch screen menu) to describe artificial intelligence in the form of his well-known and intriguing *Turing Test* for machine intelligence [Turing 1950]. This conceptual test challenges a human participant to discern, through separate anonymous dialogs with a computer and with another human being, which one is the machine and which one is the human. The Turing Test provides a compelling topic upon which to base the participation of the museum visitor.

Turing's avatar provides a narration of the exhibit and serves as an intelligent guide for the museum visitor throughout the exhibit. The visitor is first asked to select an image of one of the individuals in our research group whose photos are displayed. The photo selected by the visitor will be the basis of a second avatar to be involved in the exhibit (they are actually pre-built). Once loaded, this second avatar appears on the screen next to Turing and then progressively becomes endowed with various faculties (vision, hearing, speech and a brain) through some elementary actions by the visitor. Finally, this second avatar is subjected by the visitor to a variation of the Turing Test through a question-and-answer dialog with the visitor about a few pre-selected topics such as dinosaurs, mythical creatures and planets, either in spoken natural language or in typed text/menu choices. We consider this test to be a variation of the Turing Test because it only involves a computer (the avatar) as the contestant, rather than a human and a computer, as Turing had originally specified. (The latter would be impractical in a 5-min museum exhibit.) We consider it an extension of Turing's original test because this test employs a graphical embodiment of the computer (the avatar) that speaks in natural language and also understands natural language speech. To fool a participant into thinking it is a human on a video link, the avatar must not only be able to communicate intelligently but must also look, sound and act human-like – a more difficult condition that in Turing's original test. The conversational style of the Turing avatar's dialog is informal and encouraging, and in a language easily understood by members of the target audience. The avatars bring to life the experience of learning how artificial intelligence and computer graphics combine to build interactive virtual representations of specific humans that portray a level of intelligence.

The exhibit was formally opened to the public at the Orlando Science Center (OSC) on or about June 5, 2014, where it is currently on permanent display to the public. In the next few sections, we describe the exhibit, first in a superficial vignette as would be experienced by a hypothetical museum visitor, and then more deeply, describing the exhibit's internal workings, before discussing the results of our several evaluations. These results showed that the test subjects generally enjoyed interacting with the exhibit and their interest in computer science and artificial intelligence increased as a result of this interaction. We begin with a brief discussion of the state of the art in interactive agents.

Background and Brief State of the Art Discussion

The body of literature on intelligent virtual humans is very large. It is beyond the scope of this paper to provide an extensive review of this literature. Nevertheless, we describe some historically notable attempts to realize the dream of someday creating a virtual human equal to us in all ways except in its flesh-and-blood embodiment.

The notion of intelligent interactive agents has existed since the inception of the computing age. Idealistic visions of these agents often include extraordinary capabilities, such as seen in movies and television shows (HAL in *2001: A Space Odyssey*, Kit in *Knight Rider*, C-3PO in *Star Wars*, Commander Data in *Star Trek: the Next Generation*, and several others, most recently in the 2015 movie *Chappie*). However, in spite of millions of dollars of research over the last 50 years, state-of-the-art technology has only been able to produce a small part of these expectations.

It all started with ELIZA [Weizenbaum 1966] - the first attempt to create a conversational agent that could interact with a human user. ELIZA played the role of a Rogerian therapist with its user, and communicated through text. Despite the fact that Weizenbaum originally built ELIZA to serve as proof of the folly of artificial intelligence, the system took on a life of its own, becoming one of the first "intelligent" programs. However, there was no physical embodiment of ELIZA, and the interaction was limited, fooling unknowing users for only a few turns before becoming repetitive. Such systems later became known as *chatbots* or *chatterbots*.

Mateas [Mateas 1997] provides a comparative overview of advances in chatbots and related technologies in the 1990s. He describes systems such as the Julia project [Foner 1993] and Erin [Hayes-Roth and Doyle 1998]. Mateas, however, notes several key differences in the conversational characteristics of the chatbots of the time and the more believable recent agents - that interaction occurs in reaction to user questions in the older systems, with no regard for pursuit of a goal by the chatbot.

The 2000s presented an evolution from disembodied chatbots to embodied conversational agents, beginning with the work of Cassell et al. [2000], whose conversational playmate, Sam, gave insight into the effectiveness of human-computer interaction in a physically immersive environment. Johnson et al. [2000] discuss the concept of animated pedagogical agents (APAs) - avatars used to assist in the instruction of students. They discuss the concept of APAs, their capabilities and challenges. Bickmore and Picard [2004] presented their studies with Laura, a personal trainer agent. An early prototype of dialog-based agents, Laura's interaction with the user is one-sided and consists of question-and-answer sessions, with the Laura agent controlling the conversation.

Lee et al. 2005 experimented with using robots as conversational agents, thereby expanding the concept of embodiment to the physical world. An animatronic penguin, Mel, posed as an expert for a hypothetical product. In this work, Lee et al. supported the notion that humans could indeed interact with a physically engaging and conversationally interactive machine. This idea was further demonstrated by Kenny et al. [2007] with the Sergeant Blackwell virtual conversational agent. With a more sophisticated dialog system than Mel, Sergeant Blackwell's capabilities for conversation provide the user with a more natural human-computer interaction.

Duch et al. [2006] further expanded on the restricted naturalness evident in the template matching approach of the ELIZA-styled programs. They introduced a new concept called *concept description vectors* to implement semantic memory and increase the impression of understanding by the agent.

Further research into chatbots saw a shift towards enhanced immersive reality for dialog systems, emphasizing face-to-face avatar presentations and dialog evaluation improvements [Traum and Rickel 2002; Duch et al. 2006], including moving from text-based to speech-based interaction. Becker and Wachsmuth [2004] explored the representation and actuation of coherent emotional states in a virtual conversational agent. Kopp et al. [2005] extends this research by presenting a model for sustainable conversation in a real-world application.

Virtual humans have been used in museums before. Kopp et al. [2005] created Max, a conversational agent that serves as a museum docent that speaks face-to-face with museum visitors and provides them with general guidance about the exhibits in the computer section of the museum. Max engages the visitors in casual small talk and speaks directly to the human in natural speech, who in return, communicates with Max via typed text.

Swartout et al. [2010] created the twin avatars Ada and Grace, to serve as museum docents. They perform basically the same function as Max, except their act includes some communication among the two avatars. While they can understand spoken speech, all interaction was originally done through a human handler, who presumably knew how to phrase the question in a way that the system would understand. A more recent version [Traum et al. 2012] implements a means to directly interact with the museum visitor without a handler.

Bickmore et al. [2011] created Tinker, a cartoon-like virtual representation of a robot that also serves as a general docent. Tinker has natural speech output, and communicates directly with the audience, but through a multiple choice menu in text. Lane et al. [2013] built Mike, a virtual coach for museum visitors in the Robot Park area of a museum. It helps young visitors use software to program robots. Mike acts as an intelligent tutoring system agent rather than a docent.

Physical robots have also been used in the context of museum docents. Shiomi et al. [2006] created a robot that goes around a museum (in pairs) and help visitors understand the various exhibits therein. The robots communicate with humans via hand gestures and one-way utterances, and with each other via natural speech. Lastly, Gockley et al. [2005] created Valerie, a physical robot that serves as a receptionist in a museum to help and direct visitors with general information.

Objective of Exhibit

The main objective of this exhibit was to create an informal science education experience that could engage children of middle school age by interacting with them about computer

science and artificial intelligence in a way that was enjoyable and interesting to them. In effect, we sought to have them enjoy their interaction with the intelligent avatars, and that this enjoyment would translate into motivation for and interest in computer science (CS) and AI. While it is indeed our hope that this interest sparked in CS becomes life-long and not merely transient, we did not evaluate it for such, as a longitudinal test required to assess this was beyond the scope of our research project. We plan to properly assess this in future research. We next describe the exhibit itself.

The Turing Test Exhibit at the Orlando Science Center

In this section, we describe the Turing Test exhibit in a general and informal manner. Section 3 describes the software system that runs the exhibit as well as additional technical details behind it. We now begin by presenting the reader a technology-free vignette that describes what a museum visitor would experience. Keep in mind that the exhibit targets middle school age children and not adults, although we do expect adults as well as younger children to be occasional users.

A Vignette Describing the Exhibit

A 12-year-old child (let's call her Silvia) is visiting the Orlando Science Center while on a field trip with her middle school class. As she walks about in the museum, she passes by an exhibit about something called the "Turing Test". She doesn't know what this Turing Test is, but becomes intrigued when a British-accented voice coming from a lifelike image (*avatar*) of a well-dressed man on a screen in the exhibit kiosk beckons her out loud to help it perform this mysterious test. She becomes curious and decides to try this exhibit, pressing the bright green START button on a nearby touch screen. The exhibit then begins.

The avatar is that of none other than Dr. Alan Turing, the famed British mathematician, computer science pioneer and World War II code breaker [DiSalvo 2012]. The avatar of Dr. Turing is shown in Fig. 1 standing in front of a replica of the Colossus computer that he helped build. Turing briefly explains the test that he conceived as a way to assess the intelligence of a computer system. He explains that if one is (unknowingly) conversing with such a system but is not certain whether she/he is interacting with a computer or with a real person, then the computer system can be said to be intelligent. Turing then asks for Silvia's help in building an avatar so that she can subject it to the Turing Test. Silvia is presented with a gallery of photos of people of diverse gender, age, ethnicity and race, and Turing asks her to select one to serve as the face of the avatar to be added to the exhibit. She opts for the photograph of Myles as the face of her avatar. Upon making this selection, a pre-built avatar of Myles that has strong resemblance to the real Myles, comes on screen as shown in Fig. 2. Turing then asks Silvia to help endow the newly-created avatar with sight, hearing, a voice and a brain.

Endowing it with sight involves showing a live video of the visitor and those near him/her through a Kinect camera, with the faces outlined in colored squares that follow the motion of the subject. Endowing it with a voice means that the visitor must select from one of four synthetic voices, respectively called Crystal, Michael, Anne and Charles. The next step is to give Myles hearing, which comes after a brief explanation

Fig. 1 The Turing avatar in front of the replica of the colossus computer

of phonemes on an auxiliary screen. Finally, giving the avatar a brain involves the visitor selecting from one of five subjects that the avatar will know about: dinosaurs, mythical creatures, natural disasters, planets and the Orlando Science Center (all built a priori). Now the Myles avatar is ready to be subjected to the Turing Test by answering simple but interesting questions from Silvia in the domain selected by her. Unlike the Turing avatar, the Myles avatar is not scripted. Instead, to understand the questions, it uses an automated speech recognition system, a dialog manager, a contextual graph that holds the domain knowledge, and a text-to-speech (TTS) system to provide answers through spoken natural language.

After a few question-and-answer cycles between Myles and Silvia (limited to three question-and-answer pairs for the sake of expediency), Turing asks Silvia whether she

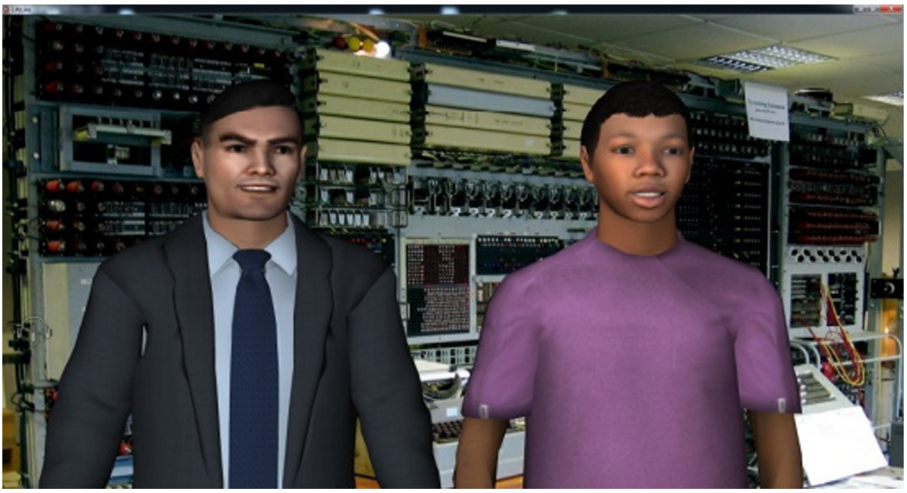

Fig. 2 Avatar of myles with turing

might be led to thinking that she was speaking to the real Myles, rather than an avatar of him - the essence of the Turing Test. In response to her answer (regardless of whether it is yes or no), the Turing avatar then states that although it is advanced, this technology still requires bright young minds like hers to help make passing the Turing test a reality for future computers. Finally, it encourages Silvia to consider a career in computer science or engineering and then says goodbye. The exhibit then automatically resets and awaits the next visitor.

The Kiosk that Houses the Exhibit

We now describe the physical implementation of the exhibit. Figure 3 shows a museum visitor interacting with the exhibit on the second floor of the Orlando Science Center. The exhibit kiosk is composed of three areas: 1) the left and right "wings"; 2) the main central section, whose frontal area constitutes the user interface; and 3) the background of the central area.

The wings of Area #1 consist of conventional textual signage describing elements of the exhibit. The left wing panel contains a brief and rather elementary discussion about artificial intelligence. The right wing panel describes the Turing Test as well as the internal components of a computer. Figure 4 depicts the wing panels. The small print on the panels is not identifiable from Fig. 4, but the figure reflects the overall appearance of the wing panels. These wing panels were not designed to be read by typical middle school students, but rather, by curious parents and/or teachers who might want to know a bit more about the exhibit and the concepts behind it as they wait while their child/children participate in the exhibit.

Area #2, the front section, is the one of most interest here, as it entails the user interface.

This section contains two screens – an upper (main) screen and a lower (auxiliary) screen. The main screen (an ASUS 24″ VH242H LCD screen) is positioned vertically at

Fig. 3 Permanent kiosk at the Orlando science center

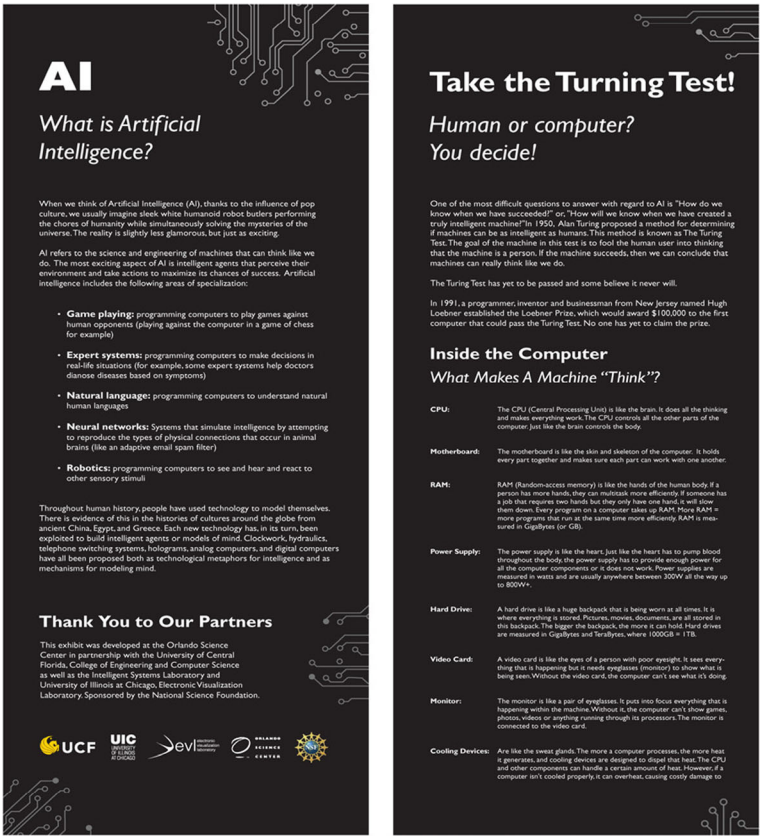

Fig. 4 The wing panels

eye level and situated inside a Plexiglas enclosure. This is where the avatars appear and interact with the visitor. All important elements of the interaction appear in this main screen. It is not a touch screen. The lower (auxiliary) screen (an Acer T321H 32″ LCD touch screen) lies nearly horizontally (with a slight tilt towards the visitor) at a waist-high level when the visitor stands in front of the exhibit. All auxiliary material appears in this auxiliary touch screen. Textual backup to the visitor is provided on this screen in case the automatic speech recognition system fails to correctly interpret the visitor's speech. Two round swivel-top bar stools are available for the visitor to sit on if he/she so wishes. They are unattached to the kiosk and can be moved around as desired. The computer that runs the exhibit was removed from its box and its components are distributed inside the Plexiglas enclosure (Area #3) for easy display. The main screen is also enclosed in this locked Plexiglas enclosure for security against theft or damage by curious visitors. Only the auxiliary screen lies outside that enclosure, given the need for physical contact with the touch screen. A virtual keyboard on the auxiliary screen is used to communicate with the avatar via text. It appears only when contextually appropriate.

The computer used employs an Intel i7 processor with 32 GB of RAM and a solid state hard drive. A GeForce GTX 670 graphics card drives the intensive graphics involved in this exhibit, with a set of Logitech X-140 speakers providing the sound. A

physical keyboard is connected to the computer, but it is not accessible to the visitor, as it is not necessary for interaction with the exhibit. It is locked inside the enclosed section of the exhibit and out of the visitor's sight. It is only used by museum staff for maintenance of the system. Finally, a directional microphone is used to capture the visitor's voice. It is located inside the enclosed section of the exhibit. A Microsoft Kinect system is used to detect motion near the exhibit. When in the stand-by mode and human activity is detected near the exhibit, it issues a loud verbal call urging the nearby person(s) to come and experience the exhibit. The Kinect web camera is also used to detect the faces of the visitor(s) when discussing giving the second avatar vision.

The Avatars

The central figure in this exhibit is the avatar of Alan Turing. The Turing avatar was created from a file photograph of Dr. Turing. While the avatar's clothing is modern (we were not able to find adequate virtual period clothing), there is a definite resemblance of the avatar to the file photograph. The dialog of the Turing avatar is heavily scripted – i.e., there are only a few situations where its reply varies in reaction to the visitor's response. Moreover, the unlikelihood that Turing's script would ever need to be changed made it feasible to pre-record the voice of the Turing avatar from a human source. The voice of Roger Thatcher, a theater student at the University of Central Florida at the time, was used as Turing's voice.

As mentioned earlier, the exhibit storyline involves a second avatar. The Turing avatar invites the visitor interacting with the exhibit to help him (it) "build" an avatar for use in the Turing Test. The visitor is implicitly led to believe that he/she is building the second avatar in real time at that moment, but such is not the case. Ten avatars, corresponding to students and faculty involved in this project, were pre-rendered and made available for selection by the visitor. The visitor is shown on the main screen the photos of the ten individuals comprising a gender, age, racial, and ethnically diverse group of people, as shown in Fig. 5. The visitor is then asked to select one for the

Fig. 5 Other avatars to be "created" by museum visitors

second avatar to be "created". Moreover, ten videos were pre-recorded - one for each of the ten individuals – that depict an artist's conception of how that specific avatar is built. The video corresponding to the selected face is played on the auxiliary screen as soon as the face selection is made by the visitor.

The avatars were built using the FaceGen Modeller system [Facegen 2012], a tool for generating 3D head and face models using front and side photographic images of the subjects to be represented by the avatars. The objective was to build lifelike avatars that closely resembled their human counterparts. The techniques used to create these avatars were developed in a prior NSF-sponsored project to create lifelike avatars. Gonzalez et al. [2013] describes the techniques used to build avatars in the "parent" NSF grant whose research results were to be communicated to public audiences through this exhibit, as part of the (henceforth discontinued) NSF CRPA program.

The human image depicted in Fig. 6 is that of Shane, one of the undergraduate researchers. Note the strong resemblance of the avatar face to the human image. Figure 7 shows the entire avatar for Shane, including the body model, as it would appear if Shane's photo was the one to be selected by the visitor.

Text-To-Speech (TTS) Voices for the Other Avatars.

Four TTS voices were purchased from http://www.nextup.com for this exhibit. Two of these are male – "Michael" (young male voice with clear American accent) and "Charles" (older male voice with slight British accent). The other two were female voices - Crystal (young female voice with American accent) and Anne (female voice with American accent). The names for the voices were arbitrarily selected and have no intrinsic meaning. The visitors have the opportunity to experiment with the voices by temporarily assigning the voices to the new second avatar without committing to their choice. They can do this for as long as they wish before committing and moving on with the exhibit. We have anecdotally noticed that most children visitors enjoy assigning a female voice to the male avatars and vice-versa. Nevertheless, most ultimately select Crystal's voice for female avatars and Michael's voice for male avatars (other than Turing, of course).

Fig. 6 Human face (Shane) and the associated avatar

Fig. 7 Full avatar for shane

Background Graphic for the Exhibit

Finding an appropriate background for the main screen that provided a suitable backdrop to the interaction was important. After considering a space station backdrop, a medieval castle and a façade of the Bletchley Park laboratory where Turing did his seminal code-breaking work during WWII, we settled on a photo of a replica of the Colossus computer, which was (arguably), his most important contribution to society. Turing begins the exhibit with a self-introduction and states that "… the electronic computer behind me is called Colossus which paved the way to smart phones, video games, and intelligent computing". This background graphic was obtained from http://wallpaper.com/wallpaper/colossus-world-382336.

Body Models for All Avatars, Including Clothing

Five body models were purchased from http://www.turbosquid.com for the different avatars. The shirts for some of the avatars were modified with the free photo editing software GIMP to change the shirt color and add a logo to the shirts. This allowed for the same body model to be used for multiple avatars without the visitor noticing that they were the same.

Automatic Speech Recognition

A commercial package was used for Automated Speech Recognition (ASR). Microsoft's Speech API, or SAPI was used as the ASR system. SAPI came integral with Windows 7, which was the current version of Windows at the time, and therefore free. The visitor's response is captured through the directional microphone mentioned above. The microphone is turned on when the avatar that is speaking stops speaking and requires a response from the visitor. The microphone automatically turns off after the ASR system provides the text of what has been uttered. However, after two failed successive speech recognition events, the system assumes that there are environmental issues affecting the recognition, and forces the user to use the touch screen on the

auxiliary screen for providing input. A failed speech recognition event occurs when the ASR returns text that does not allow the dialog system to make a selection for the question being asked by the visitor. It is also considered a failed speech recognition event if the visitor does not provide a response that the system is expecting from the context of the dialog. Microphone de-activation is indicated by a small microphone icon at the bottom left of the main screen that turns yellow upon de-activation - a change from its normal gray color. The visitor can always opt to directly interact with the exhibit via the touch screen on the auxiliary screen, and many do. Chant Speechkit V5 is used to interface with the Windows 7 SAPI ASR. Chant allows the exhibit to simply request the microphone be turned on and retrieve what was recognized. The Chant software can be found at http://www.chant.net/Products/SpeechKit.

Other Graphics

There are several other graphics used to make the exhibit visual to the greatest extent reasonable. These are too many to describe here and most of them are inconsequential. Nevertheless, Fig. 8 shows the touch sensitive graphic used to permit a selection of the topic about which the second avatar is to converse while subjected to the Turing Test. This selection appears on the auxiliary screen when the second avatar is to be endowed with a brain.

The Exhibit Framework

In this section we provide a technical description of the exhibit. We begin with the *Exhibit Framework* (EF), the software system that controls the flow of the exhibit. It provides the infrastructure to implement the exhibit and controls all interactions with the visitors. The Exhibit Framework (hereinafter simply called EF or simply the *framework*) is the most complex component of the exhibit. While all other components described above were acquired commercially and the kiosk was built by the Orlando Science Center, the Exhibit Framework was entirely built in-house by our research team. We believe that our Exhibit Framework is flexible and general for use in other exhibits that involve interaction with avatars. However, not having specifically evaluated this feature, we make no claims about its validity here. We hope to address this issue in future research. This section and its subsections describe the Exhibit Framework.

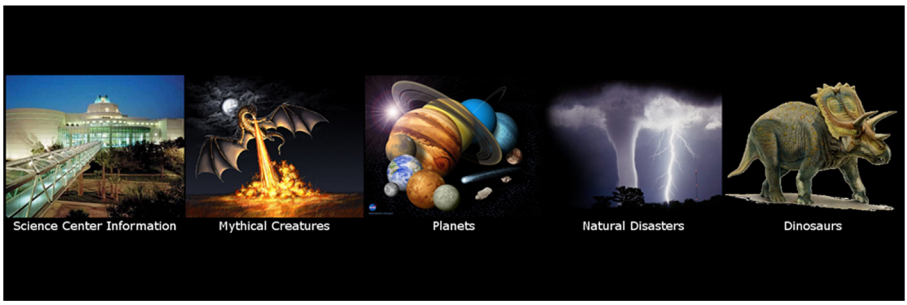

Fig. 8 Icons appearing on the auxiliary screen indicating the domains of knowledge

Contextual Graphs

At its most fundamental level, the Exhibit Framework uses a modified version of Brézillon's *Contextual Graph* (CxG) paradigm [Brézillon 2003] to hold the knowledge required to run the exhibit. CxGs are acyclical directional graphs that ask a visitor (or an intelligent system) to further refine the context of an inquiry by answering questions or selecting alternatives. CxGs represent procedural and declarative knowledge, extending traditional graph-based knowledge representation strategies such as Semantic Networks, with roots in marker-based reasoning paradigms which utilize them [DeMara and Moldovan 1991]. The EF system executes the graph, thereby providing the "motion" of traversing a CxG from its beginning at its single entry point, to its single exit point at its end. A brief description of CxGs follows. See Brézillon 2003 for a detailed description.

A CxG organizes the knowledge in a way that always seeks to provide an unambiguous answer/response/decision/solution to a question/problem/inquiry that may be ambiguous. It does so by asking questions whose answers will eliminate the ambiguity. This obviates the need for managing uncertainty. To accomplish this, it continually seeks user input to progressively refine the context in as fine-grained a manner as possible to eliminate the uncertainty. Once the context is defined at a sufficiently fine grain, the answer becomes clear and unambiguous. While this concept does not work for problems where questions may not have definitive answers, it does work well in our exhibit because all questions therein do indeed have clear and specific answers.

A CxG is composed of context-sensitive *activities* or *actions* whose execution can be triggered by the answers provided to queries. In a "standard" CxG (as defined by Brezillon), there are five basic components in a decision tree-like contextual graph. A generic standard CxG that contains all of these possible elements of a CxG can be seen in Fig. 9 [Brézillon 2003]. These elements are shown in boxes or circles of different colors and shapes, with numbers assigned to them. These numbers located inside each element have no intrinsic meaning and are used solely to identify the particular component here. The first type of component described is the *Contextual Node.* It is represented in Fig. 9 by the large (blue) circles numbered 1, 2 and 13, and are located on the left side and the lower part of the graph. The Contextual Node represents a decision that must be made during the execution of the contextual graph to refine the current context. More practically, it dictates which branch of the CxG is to be taken next.

The second type of component in a CxG is the *Action Node*, which is denoted by the (green) squares numbered 3, 7, 8, 10, 11, 12, 14, 15, 16, 17, 18 and 19. The Action

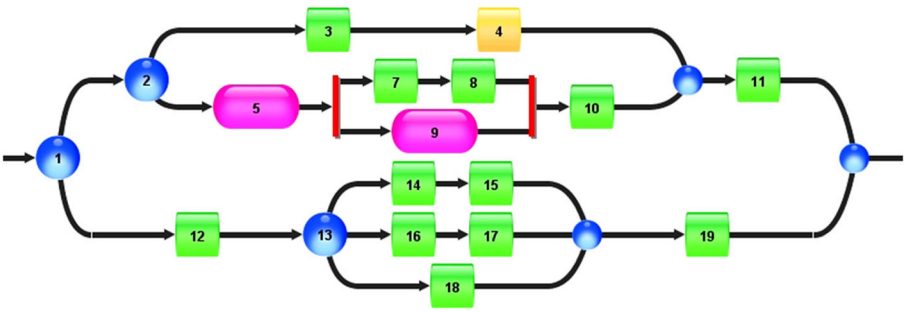

Fig. 9 An example of a generic contextual graph [Brézillon 2003]

Nodes represent actions that must be completed in order to achieve the desired outcome. The third type of component, an *Activity Node*, is represented by the (purple) ovals numbered 5 and 9. An Activity Node is an embedded contextual graph - it abstracts and represents an entire contextual sub-graph within the larger CxG. It can be expanded to display the details of this sub-graph. The fourth type of component of a "standard" CxG is a *Gap*, which is denoted by the (red) thick vertical lines before and after Action Nodes 7 and 8 in Fig. 9. A Gap signifies that all of the paths encompassed by it must be taken, but the order of which paths to take first is arbitrary and not important. The fifth and last type of component is represented by unnumbered smaller (blue) circles called the *Recombination Nodes*. They have no intrinsic function other than to recombine paths and direct them towards a single exit point. Thus, they are considered to be inert. The arcs in the graphs (black lines) indicate paths in the contextual graph.

Our Exhibit Framework uses three of the components of the standard contextual graph (Contextual Nodes, Action Nodes and recombination Nodes) and introduces a sixth one, the *Goto* node represented by the (orange) square box (#4) in Fig. 9. The Goto node allows the exhibit control to change contexts quickly when required by the evolving situation with the user. The addition of the Goto node now makes the CxG potentially cyclic. However, this particular exhibit does not make use of Goto nodes.

How CxGs Are Used by the Exhibit Framework to Control the Exhibit Flow

The framework reads in a CxG file and proceeds to execute it, thereby providing progression through the exhibit. In our modified version of the original CxG paradigm, the Action Nodes implement everything that is said by the avatars and that is displayed or done on both screens of the exhibit. For example, everything that Turing could say during the exhibit and all of the different graphics and videos displayed on the auxiliary touch screen are stored in several different Action Nodes throughout the main control CxG (referred to as the *Main CxG*). The Action Nodes also allow certain internal variables (flags) to be set at the beginning of the exhibit to be accessed later. Decisions based upon these internal variables and responses from the users are handled by the Contextual Nodes in the CxG. This allows the framework to know and use the current context. For example, the second avatar should not speak until the visitor has given it a voice. Secondly, the newly created second avatar should try out its new face by making grimaces immediately upon its initial appearance on screen. The EF system keeps track of the visitor's inputs and can direct the next actions to be taken by the exhibit accordingly.

The full Main CxG for the Turing Test exhibit is shown in Fig. 10. Although it is unlabeled and impossible to discern its details, we include it to display its overall magnitude. The points where a path splits into several paths are the Contextual Nodes, while the small circles that resemble dots along the path are the Action Nodes. All these nodes have significant content but it is not possible to include these in this Fig. A CxG defines the current context as the current position in the graph as it is being traversed. Our exhibit design requires that all possible decisions be pre-defined and pre-programmed. While this could be considered a limiting factor in how an exhibit is built, we did not find it to be so during creation of this exhibit because the main control of the exhibit was predominantly pre-scripted anyway, with only a relatively few decisions necessary along the way. These possible decisions are depicted in the alternative paths emanating from a contextual node. It can be seen from Fig. 10 that creating this exhibit required significant effort.

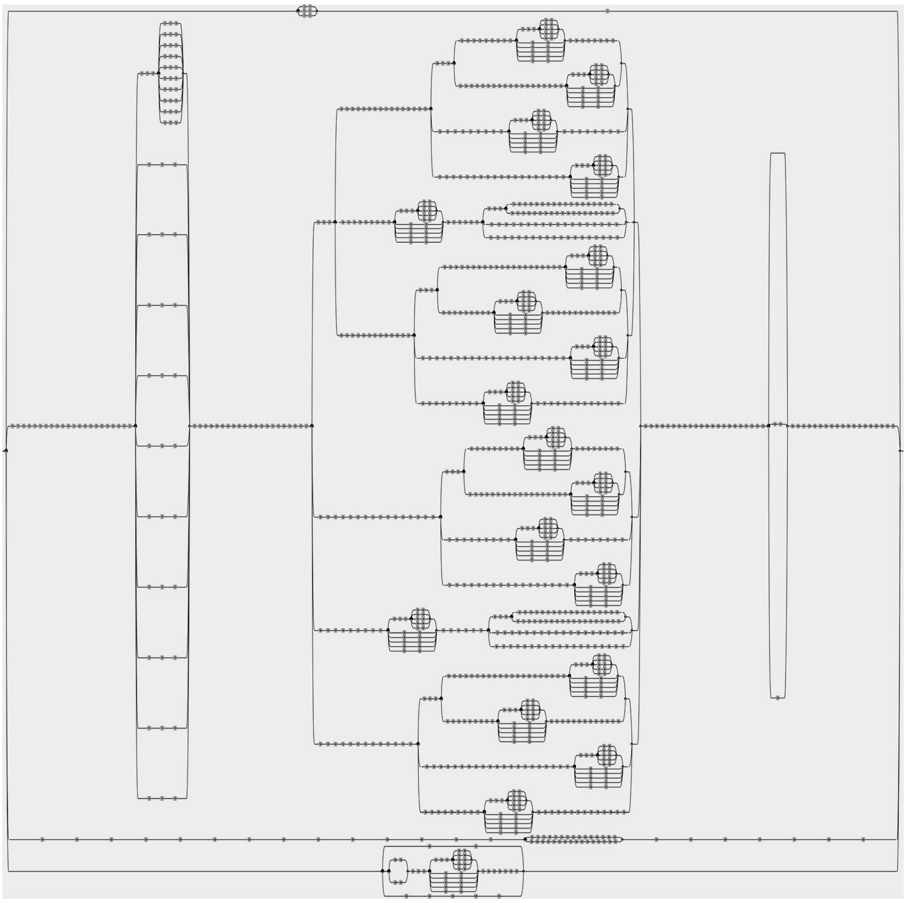

Fig. 10 Full main CxG used to control the exhibit

There are a total of **1024** nodes in the Main CxG that controls the exhibit, with **963** being Action Nodes and the rest being Contextual Nodes. Because of their inert nature, Recombination nodes are not counted. All of these are reflected in Fig. 10.

Each Action Node contains a different command, plus arguments for that command. An example of the content of an Action Node for a "SAY" command is "SAY//TUR//I am Dr. Alan Turing. Years ago I helped create ……" The first item identifies the command. "SAY" is a command for an avatar to speak something. Another command labeled "LOD" loads an item onto one of the screens, most often to display an image file. "MOV" commands an avatar to move to a different position on the screen and "EXP" commands an avatar to display a different facial expression. The rest of the contents of each Action node (its arguments) are based upon the type of command it holds. For the commands that affect an avatar (SAY, MOV, and EXP), the second item after the command identifies to which avatar the command applies. In the example of the SAY command above, TUR is the Turing avatar. The last part of the content of the action node in the example above includes the exact words for the Turing avatar to utter through TTS. We should point out that this example is for illustrative purposes only, as in this exhibit the Turing avatar doesn't use TTS but rather, pre-recorded human voice. In this case, the Main CxG actually uses

SAY commands to dictate what Turing should say. The text that Turing is to utter is compared to text in an XML file that indicates which audio file to execute if a match is found between the text specified by the Main CxG and the text in the XML file. If the exhibit developer instead desired to generate the Turing avatar's speech through TTS, the EF system would easily accommodate this.

The Contextual Nodes are decision points in the CxG where a visitor's interactive input is sought and obtained. The input received then dictates which of multiple alternate paths in the CxG is to be followed in the progression through the exhibit. For example, when the progress of the exhibit reaches the point where visitor must select the face of the second avatar, the relevant contextual node in Main CxG will ask the visitor which face he/she would like to select for the avatar's appearance. The visitor can respond by simply saying the name of the avatar ("James"), or utter complete sentences such as "I want James". Alternatively, the visitor can simply touch the desired photo shown on the auxiliary screen (see Fig. 5). This causes a specific segment of the path in the CxG to be identified, and the course of the exhibit then continues down this path. Note that there is a timer that repeats the last question if user input is not received within a specified period of time. After a question is repeated four times without a response, the Exhibit Framework assumes that the visitor has left and resets the exhibit for the next visitor.

The control of the exhibit is done via the Exhibit Framework system as it executes the Main CxG and implements the commands and decisions in its nodes and paths as it comes upon them. The EF system creates a list of Action Nodes whose embedded commands are to control everything that occurs in the exhibit, from what each avatar says to where they look and move. When the EF system initially loads, the contents of the Main CxG are read and a list of Action Nodes containing node numbers is created. Each node in the CxG is assigned a unique number that corresponds with its position in the list. Upon pressing the start button, the EF system sets an internal mode variable to "exhibit". The EF system always begins at node 0 of the Main CxG, which is a Contextual Node that checks this internal mode variable. If set to "exhibit", the EF system begins to traverse the graph. This involves extracting commands from the Action Nodes in the aforementioned list and populating a command queue with this sequence of commands, until a Contextual Node is reached whose question has to be answered by the visitor at an appropriate time. Once the last command has been processed and the progression reaches the next Contextual Node, the system asks the relevant question in the Contextual Node encountered, waits for an answer, and then processes the visitor's response to determine which path to next take in the graph. Upon receipt and interpretation of the visitor's input to the question asked, the node list is repopulated with the nodes in the path extension up to the next Contextual Node. The command queue is then built again from the Action Nodes in the new path and the commands are sequentially executed. This process is repeated until the end of the CxG is reached, signaling the end of the exhibit.

The Dialog Management System for Avatar Q&A

A context-based dialog management system uses CxG-based knowledge bases to respond to questions in each of the five different domains known by the second avatars. Note that the CxGs used by the dialog manager in the Q&A session are separate and different from the Main exhibit CxG described above, and are only used in the question-and-answer

interaction at the end of the exhibit. This is when the visitor engages the second avatar in a Q&A session on the selected topic as part of the Turing Test.

Our dialog management system is a variation of Hung's *CONtext-centric Corpus-based Utterance Robustness* (CONCUR) system [Hung and Gonzalez 2013]. During the early stages of the CONCUR platform's development, it was found by its developers that in their specific application, their ASR system had a very high word error rate (WER), one which would normally preclude effective speech-based interaction. We define WER as:

$$\mathrm{WER} = (\mathrm{S} + \mathrm{I} + \mathrm{D})/\mathrm{N}$$

where I = number of words missed by the ASR system that needed to be inserted; D = number of superfluous words put out by the ASR system that needed to be deleted, S = Number of incorrect words put out by the ASR system that had to be directly substituted by correct words, N = Total number of words in the reference string.

This led the CONCUR developers to implement techniques that would lessen (although not totally eliminate) the negative impact of high WERs. To accomplish this, CONCUR uses a lightly annotated corpus (knowledge base) that is contextually organized. These corpus items are hierarchically structured as topics and sub-topics. This organization allows the system to immediately categorize the knowledge into different contexts. These contextualized corpora are further processed into a conceptual signature, consisting of a set of characteristic key phrases. This parsing and tagging is performed by the WordNet-enabled C# library named SharpNLP, which is based on the OpenNLP open source project. The extraction process entails tokenizing the input sentence into phrases - CONCUR pays special attention to the noun and verb phrases. Similarly, when a user's utterance is received during a typical interaction, the input text obtained by the ASR system is also processed through the key phrase extractor.

All free-form responses from a user (text or spoken, but NOT the touch screen menu selections) are first tagged by their part of speech with SharpNLP. The nouns become the keywords used to determine the appropriate path in the graph. Using the user's extracted key phrases, CONCUR performs a two-layer context matching search in the knowledge base. The initial search layer looks for an exact match - does what the user said match a contextual topic word for word? This exact search is fast and has the potential of finding a suitable context in the corpus, but it has a low success rate, as most people tend not to mention the contextual topic verbatim in a conversation.

If the verbatim search fails, a second search layer is summoned to compensate for the variations in word selection of the user, along with any errors introduced by ASR. During this search, each extracted utterance key phrase is compared to the different contextual signatures in every corpus, looking for partial matches. The matching process for this search utilizes a WordNet-based conceptual matching system. Under such a system, conceptual similarity comes into play, allowing for synonyms and word order independence.

When a partial match is found with a contextual topic, that topic is added to a list of possible context matches. After all the contextual topics in the knowledge base have been evaluated, the system analyzes this list of context possibilities. If this list only contains one item, an answer prescribed within that context is presented to the user. However, if the list contains multiple possible matches, then the user is asked to clarify what he/she asked. CONCUR always presents the first match, and the matching process

is performed again on this subset of contexts. If no matches are found, it continues on with the previously-determined context.

In the case where a context match is satisfactory, CONCUR pulls several different sentences from the corpus of the topic identified by the context, and in the order they appear in the text, and assumes that these sentences will answer the question. This means CONCUR can only be used to answer general questions ("Tell me about life on Mars") and not specific questions ("When did Apollo 11 land on the moon?"). This inability to customize a response or interpret information from the corpus represents the main limitation of CONCUR.

To address this limitation for this exhibit, CONCUR was modified by replacing the annotated corpus normally used with CONCUR with a contextual graph. This would permit some reasoning about the knowledge to be given in response to a question by the user. All of the possible responses by the avatar are stored within their own individual Action Nodes, and each path through the CxG will contain only one Action Node. This allows the system to stop processing as soon as the Action Node is located on a path. The Contextual Nodes are used to refine the context of each of the responses and thus determine the path explored by the dialog manager while seeking to answer a question. For example, the knowledge base includes information about the different planets. Figure 11 shows part of the CxG for the planets knowledge base. The first Contextual Node (#1 in Fig. 11 – the first large (blue) circle on the far left of Fig. 11) would contain the names of the planets as the broadest of topics. It would identify the name of the planet from the user's utterance.

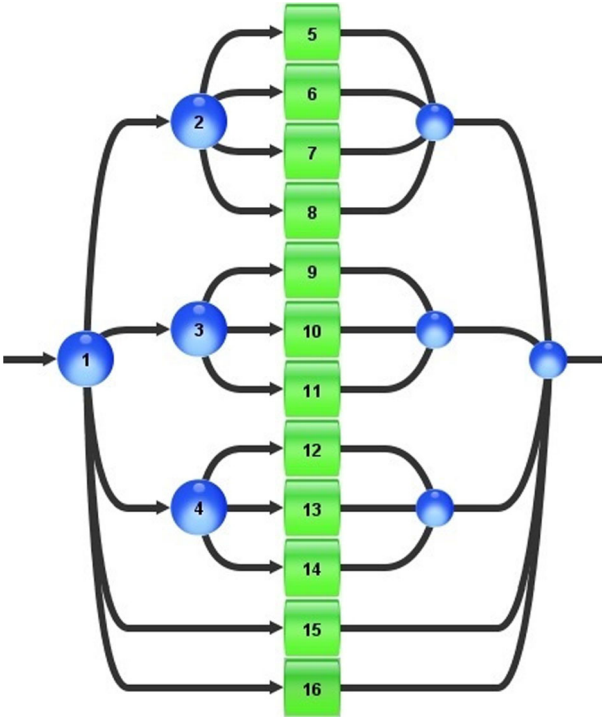

Fig. 11 An example of the knowledge base format

The second set of Contextual Nodes (#2, #3, and #4 - the three large (blue) circles on the left side) in Fig. 11 - contains keywords that help the system refine the context in order to find the correct response for the question asked. In our example, this second level would contain general information such as the size of the planet, its composition (gas, rock), its rotational cycle, etc., along with any common synonyms for those terms that may be included in the utterance. Once this has been answered, the CxG continues to the appropriate Action Node (the square (green) boxes #5 through #16 in the middle of the graph), which contain the answers to be provided by the avatar.

The overall dialog manager search algorithm is shown in Fig. 12. The same **Keyword (key phrase) Algorithm** from CONCUR is used in the modified dialog manager.

The knowledge base representation paradigm is not the only change made to CONCUR. Another modification was in the way the searches are conducted. The exact search was eliminated, as it was found to be highly atypical for this search to find a context, let alone the correct sub-context. The system now begins with the second keyword search from CONCUR, the **new search** (see Fig. 12). In the improbable event that this search fails to find the context of what the user asked, a second search called the ***contains search*** was added**,** where the system looks for keywords within words (partial string matches) to help identify the proper context. For example, if the only keyword listed to end the conversation was "bye" and the speech recognition detected "goodbye," then the **contains search** would recognize that "goodbye" contains "bye" and select that context. For the **contains search**, words with two letters or less are neglected. This prevents the system from matching common words such as "is" and "a" to the improper context. The **contains search** ensures that if there is any possibility of finding the correct context, it should be able to be found unless the speech recognition completely failed to recognize any words at all in the utterance, which is unlikely.

To better understand how this functions, we now walk through a plausible interaction using the planets knowledge base. Before the interaction occurs, the system reads into memory the CxG containing the knowledge base selected by the visitor when "giving" the avatar a brain; in this example, "Planets" was selected

1. Process visitor's utterance using **Keyword Algorithm**
    a. Tag user's utterance
    b. Remove everything except for nouns
    c. Keywords are the nouns
2. Check if the list of possible matches is empty
    a. If not empty, **Previous Search Algorithm**
        i. Compare new keywords to matches found in last search
3. Check if the list of possible matches is empty
    a. If empty, **New Search Algorithm**
        i. Look for matches in knowledge base using exact keywords
4. Check if the list of possible matches is empty
    a. If empty, **Contains Search Algorithm**
        i. Look for matches in knowledge base using words within words search
5. Check the list for possible matches
    a. If no matches, return a don't understand
    b. If 1 match, return match as right response
    c. If multiple matches, find common keyword among matches
        i. Inquire further about common keyword

Fig. 12 Search algorithm for the dialog manager

by the visitor. This section of the exhibit begins with the second avatar asking, "What would you like to know about Planets?". A list of five possible touch-sensitive questions (i.e., contexts) are randomly selected and displayed on the auxiliary screen to give an undecided visitor some suggestions of what to ask. For this example, the visitor chooses to say "Tell me about Mars", rather than touch one of the five suggested questions. Assume then that the speech recognition system then only detects the word "Mars" in the visitor's spoken request. After completing the **new search** algorithm, the system will have found several matches, which causes it to skip the **contains search**. The dialog system analyzes the several matches and determines that most of these possible matches are within the Mars context. All the possible matches that are not within the Mars context are removed from the list. The second avatar then inquires what the visitor would like to know about Mars. On the touch screen, the visitor is now shown up to five of the possible matches - fewer if the possible match list is smaller. The visitor's next utterance asks whether there is life on Mars. The dialog manager now compares the words detected by the ASR against all the sub-contexts within the Mars context using the ***previous search*** algorithm (see Fig. 12). If a match is found (e.g., "life"), the found response about life on Mars is related to the visitor as Text-to-Speech (TTS) (e.g., "It is strongly believed that there is currently no life on Mars"). If no matches can be found within the current context, the system assumes that the visitor may want to change the current context and commences the **new search** algorithm from the beginning.

After the visitor asks three questions, the second avatar is deactivated and the Turing Avatar once again takes control of the dialog, asking whether the visitor might have thought he/she were speaking to a real person via a teleconferencing system. Regardless of the visitor's response (Yes or No), it will go on to say that the technology is not yet mature enough for a meaningful Turing Test and exhorts the visitor to consider helping it to achieve this by becoming a computer scientist or engineer. Lastly, Turing says thank you for helping it out and good bye. This exits the Main CxG and concludes the exhibit.

Assessment

This section discusses the assessment of our work. In particular, we sought to determine how well the exhibit accomplished its stated objective - the engagement achieved by the museum visitors who interact with the exhibit. This was measured indirectly by asking: a) how did the visitor enjoy the exhibit; and b) how much did the exhibit serve to increase the interest of the visitor in computer science/engineering. There were two types of tests performed: *Formative testing* with exhibit prototypes during the development of the exhibit, mostly in backrooms and with test subjects selected from the museum floor; and *summative testing* performed with the final delivered system after the final version of the exhibit had been open to the public on the museum floor. Formative assessment is normally done during the course of the development to provide feedback to the development team. This feedback can be incorporated as improvements in the next iteration of the product design. A summative assessment serves as a final judgment of the project and how well it met its stated objectives. We begin with a description of the formative assessment.

Formative Evaluation of the Exhibit Prototype with Museum Audiences

We built a series of successive prototypes and subjected these progressively-improved prototypes to formative evaluations in the hands of museum visitors at the Orlando Science Center. While these evaluations asked several questions related to different aspects of the exhibit on which feedback was desired, the discussion here focuses on how well the objectives of the project were being met. Improvement in meeting these objectives was used as a measure of the effectiveness of this formative feedback. There were three formative evaluations; the first two were held in a museum backroom using test subjects selected from the museum floor by museum personnel, while the last was held on the museum floor and open to all visitors.

Formative Evaluation #1 – Back Room Testing with Prototype System on a Table:

Museum visitors were selected by OSC personnel from the main floor of the museum and asked whether they would be willing to participate in a research project. Upon agreement, they were taken to a back room and were allowed to operate and experience a prototype version of the exhibit without any help or interruption. An age-diverse group was sought in spite of the fact that our target age was middle school children. This was done to increase the size of the pool as well as to gauge the differences in opinion by members of different age groups. The visitor was asked his/her education level during a post experience interview, and their answer was used as indicator of the age of the subject in the analysis. No pre-interview was carried out.

The evaluation instruments consisted of the following:

- The visitors' interaction with the exhibit was observed by members of the team as well as employees of the OSC. The duration of the interaction was logged, as an excessively long duration of the exhibit was one of our major early concerns.
- At the end of the interaction, the test subject was interviewed by a member of the research team, and a survey was completed by the latter who recorded the responses provided orally by the test subject. A Likert scale of 1 to 5 was used, where 1 was a highly negative response and 5 was highly positive. We should note that the test subjects did not see the survey itself. The interviewing team member asked the questions and recorded the answers. In the case of questions that were phrased as binary, the interviewer asked for a quantitative assessment in the scale of 1 to 5 and explained what the various responses would mean. This was done because many of the test subjects were children and we sought to maximize the probability that these underage test subjects could understand the question being asked and would be able to answer them in a manner reasonably consistent with those of the other test subjects.

The Appendix includes the survey used in these tests. Table 1 summarizes the responses to the most important questions – Q11 ("Did the exhibit increase your interest in computer science/engineering?") and Q12 ("Did you enjoy the exhibit?"). In Question 12, we implicitly equated exhibit enjoyment to engagement, as directly asking children about engagement (e.g., "Did you find the exhibit engaging?") would have surely seemed confusing to them, thereby distorting the interview process. The rest of the questions (Q1 through Q10) involved issues that were important at the time in order to design and

Table 1 Results for Q11 and Q12 in the first formative assessment

Measurement	Q11 - Did the exhibit increase your interest in Computer Science/Engineering?	Q12 - Did you enjoy the exhibit?
n	48	56
Mean	3.16	3.92
Standard Dev	1.42	0.77

maintain the exhibit, but are not directly relevant to this paper. We should note here that Questions #1 to #10 were somewhat different in the three surveys used for the different formative evaluations. We show the one used for the first formative assessment. Nevertheless, the wording of questions #11 and #12 was identical in all surveys.

The results were deemed generally positive in this round of tests, where 3.0 was seen as a neutral mid-point in the range of answers. We interpreted low numbers for increased interest (Q11) to mean absence of increased interest, rather than a decrease in interest. However, some interesting insights can be gleaned from the results of Table 1.

- The number of respondents for Q11 was eight fewer than for Q12. For some reason, some subjects refused to answer Q11. We did not formally explore the reasons for this, but believe that children who did not know how to answer this question simply opted to not do so.
- Enjoyment (engagement), as indicated by Q12, was scored higher than interest (Q11) in most responses. We likewise did not formally seek to explain this, but believe that the reason is that children more readily recognize their own enjoyment of something than their interest in same.

The duration of the exhibit (approximately six minutes on average) was nearly unanimously considered about right by the test subjects across the age spectrum. The response averaged 2.98 on the Likert scale (see Table 6 below). By this we mean that a response of or near 3.0 on the Likert scale would be ideal – not too long and not too short. This was a highly positive result. However, we still sought to reduce its duration further (to approximately 5 min) in spite of this result, as we were not convinced that the duration was not too long.

Formative Evaluation #2 – Back Room Testing with Prototype System in Temporary Kiosk

The same procedure used in Evaluation #1 was followed in this second set of tests approximately six weeks later. The difference was that the prototype used was modified to incorporate the feedback received as a result of the first formative evaluation. This included a slight shortening of the exhibit and clarification of several ambiguous statements by Turing and the other avatars. The same Questions #11 and #12 contained in the Appendix were used, and the process for gathering the data was likewise the same. One difference was that rather than simply placing the computer on a table as done in the first evaluation, a rough prototype kiosk was built by OSC and used for these tests. However, no questions in the survey asked about the kiosk. A summary of those results are shown here as Tables 2 and 3. These tables only contain the results for Q11 and Q12 – the only two questions deemed relevant to this paper.

Table 2 Q11: Did the exhibit increase your interest in Computer Science/Engineering?

Measurement	Elementary School	Middle School	High School	College	Post College	Overall
n	5	11	25	0	15	56
Mean	4.80	4.18	3.52	-	3.7	3.83
Standard Dev	0.45	1.47	1.08	-	1.44	1.26

The results were substantially more positive than for Formative Evaluation #1. The results for enjoyment increased, especially among the target age group. It also increased among the elementary school children. This argues the point that this age group might also be included in the target age group, as they are likely to enjoy interacting with animated characters (the avatars), even if some of the younger ones may not totally understand the concept of the Turing Test.

More encouraging was the increase in declared interest in computer science/engineering expressed by the participants. This significantly increased in the target age group to strongly positive, as well as mildly positive for high school students. This was an improvement for the stated project objective of stimulating interest in computer science among middle school students. The results for the other 10 questions were equally positive, making this round of tests quite successful.

Formative Evaluation #3 - Museum Floor Testing with Prototype System in Kiosk:

At this point, the research team felt confident that no additional formative tests under backroom conditions were necessary. We proceeded to carry out the next set of tests under more natural conditions: setting the exhibit in the museum floor and waiting for visitors to naturally approach it and experience it. Visitors that did experience it in its entirety were approached afterwards and asked whether they could be interviewed, using the same questions #11 and #12 used before in Formative Evaluations #1 and #2. Tables 4 and 5 summarize the results obtained. The same kiosk used in Test #2 was used here.

The first location selected by the OSC was immediately outside the entrance to a theater where periodic showings of a movie were held. This location was outside of the main stream of traffic in the museum. As a result, only a few visitors experienced the exhibit in the first few hours of testing. Several visitors on the way to the theater for a showing approached the exhibit and hit the start button, but because of the pending start of the showing at the theater, did not have the time to fully experience the exhibit and prematurely abandoned it sometime during its operation, often at the urging of their teacher and/or parent. After conferring with OSC staff members, it was decided to move it to a more crowded place where it would have a better chance of attracting visitors. The new location

Table 3 Q12: Did you enjoy the exhibit?

Measurement	Elementary School	Middle School	High School	College	Post College	Overall
n	5	11	25	0	15	56
Mean	5.0	4.90	4.36	-	4.06	4.44
Standard Dev	0.0	0.30	0.86	-	0.96	0.82

Table 4 Q1: Did the exhibit increase your interest in Computer Science/Engineering?

Measurement	Elementary School	Middle School	High School	College	Post College	Overall
n	2	1	2	1	2	8
Mean	4.50	4.00	4.00	4.00	5.00	4.38
Standard Dev	0.71	N/A	1.41	N/A	0.00	0.74

resulted in several more visits, although because of the lateness of this shift of location, the number of museum visitors using the exhibit was low, and thus, a sufficiently large data set was not captured. Therefore, reliable claims cannot be made from Formative Evaluation #3. Nevertheless, being on the museum floor, this was considered a more realistic test of the exhibit's effectiveness, and the general consistency of the results with Evaluations #1 and #2 gives us some confidence that these results are valid.

Summary of Formative Evaluations

Table 6 shows the compiled results for the three formative evaluations. This table provides a clear overview of the results, although it is not broken down by age or gender. The numbers for the two most important questions: interest in computer science/engineering and whether they enjoyed the exhibit, were found to progressively increase in the three formative evaluations. Particularly encouraging was that the mean for all participants in Q11 ("Did the exhibit increase your interest in CS/Engr?") went from **3.16** in the first evaluation, to **3.83** in the second evaluation, to **4.38** in the last one. This is a notable improvement in the results, and indicative of the promised impact of this exhibit in young visitors pursuing STEM careers. These encouraging results must be tempered by the fact that the number of test subjects for the third set of tests was too low to make a statistically valid assertion. Further tempering the result is that this is only one question, and it is self-reported, with all the accompanying implications that self-reporting has. As mentioned in Section 1.2, it is our hope that whatever interest sparked in CS/Engr is life-long and not merely transient; however, we did not evaluate it for such, as a longitudinal test was beyond the scope of our work.

Likewise, the mean of Q12, ("Did you enjoy the exhibit?") went from **3.92** to **4.44** to **4.56**. One must note once again, that the low number of test subjects (9) in the last evaluation makes this conclusion only tentative. However, the results trend in the correct direction, in spite of the museum floor environment being more likely to distract the test subjects.

The perceived duration of the exhibit hovered around 3 in all three evaluations. The average of averages, weighted to compensate for sample size, was **3.03**. This indicates that the duration of the exhibit was deemed just right - neither too short nor too long.

Some of the results for other (i.e., "Were you able to understand Turing?", "Were you able to understand the second avatar?") were somewhat lower for the third set of results.

Table 5 Q12: Did you enjoy the exhibit?

Measurement	Elementary School	Middle School	High School	College	Post College	Overall
n	3	1	2	1	2	9
Mean	5.00	5.00	4.5	5.00	3.5	4.56
Standard Dev	0.00	N/A	0.71	N/A	2.12	1.01

Table 6 Summary of results for the three formative evaluations

Date of Test ➔ Question V	Feb. 23/24, 2013		April 2/7, 2013		August 31, 2013	
	Mean	StdDev	Mean	StdDev	Mean	StdDev
Q1 – Age (school grade)	-	-	-	-	-	-
Q2 - Gender	-	-	-	-	-	-
Q3 – Was it too long?	2.98	0.46	3.11	0.49	2.86	1.23
Q4 - Enjoyed Eliza interaction?	3.19	1.24	3.27	1.17	-	-
Q5 – Understood Turing?	4.44	0.70	4.53	0.68	4.44	1.01
Q6 –Understood other avatars?	3.91	0.94	4.28	0.77	3.89	1.27
Q7 – Able to follow exhibit?	4.06	0.94	4.24	0.83	4.63	1.06
Q8 – Notice Background?	-	-	-	-	-	-
Q9 – Flowed smoothly?	3.82	0.91	4.20	0.92	4.63	1.06
Q10 – Changes? (open ended)	-	-	-	-	-	-
Q11 – Increase interest in CS/E?	3.16	1.42	3.83	1.26	4.38	0.74
Q12 – Did you enjoy exhibit?	3.92	0.77	4.44	0.82	4.56	1.01
Number of participants (Q11/Q12)	48/56		56/56		8/9	

This was expected, as the exhibit was now placed in a noisy museum floor, rather than in a quiet back room. Nevertheless, the results obtained for these were deemed satisfactory.

Interestingly, in the last two formative evaluations, data suggests that the exhibit was received better by elementary school aged children than middle school aged ones. This was somewhat surprising initially. However, given the general optimism and willingness to please of young children, maybe it shouldn't be surprising. We will further investigate this finding in future research.

Another note of interest was that the responses for enjoyment of the exhibit were generally more positive than those about increased interest in STEM. To investigate this issue further, we sought to determine whether there was a statistical correlation between enjoyment of the exhibit and increased interest in STEM. Given that the responses were in a Likert scale format, we applied the Kendall rank correspondence test using the Gamma metric. The result was a correlation factor of 0.4967, indicating mild correlation. To confirm, the Spearman Rank Correlation test was also performed, and obtained a correlation factor of 0.442 and a p-value <0.01. This test also suggests a mild correlation between the two responses.

Lastly, the results for the second and third evaluations were generally higher than for their previous evaluation (i.e., #2 better than #1 and #3 better than #2). This is a testament to the value of formative testing and how incorporating the feedback obtained from the test subjects improved the exhibit.

All final suggestions were incorporated into the exhibit and the final version of the exhibit was deployed on the museum floor on or about June 5, 2014. The last remaining assessment was now the final, summative assessment. We discuss this next.

Summative Evaluation

The summative evaluation was completed on July 31, 2014, after the exhibit had been formally opened to the public in its final form for almost two months. The research team of four (the PI and three students) set up a table approximately 100 ft away from the exhibit – far enough away that the exhibit visitors were not aware that they were being observed. To further disguise the true objective of the table and the people sitting

behind it, the table was made to have the appearance of a general help desk. Individuals and parties visiting and engaging with the exhibit were observed and notes were taken. No electronic recordings of the interactions by the museum visitors were made because of privacy concerns in test subjects who would be unaware of their roles.

Data collected included the time spent interacting with the exhibit; the level of engagement with the exhibit; the approximate age of the main participant (estimated by school grade); the number of participants in the party; and the likely reason for abandoning the interaction if the experience ended prematurely. The most important data taken were the duration of the visit and whether they saw the exhibit to its completion. Also observed and judged was the level of engagement with the exhibit. While this was a qualitative value judgment made by the observers, the intensity of the visitors while experiencing the exhibit was noted and classified by vote of the four observers. Three levels of engagement noted were high *engagement, medium engagement,* and *low engagement*.

To eliminate inclusion of incomplete data, groups that did not have both, the level of engagement and the duration of the exhibit interaction recorded were excluded from the analysis. This reduced the total sample size from 42 to 38 sample groups used for the analysis.

Definition of Engagement Levels

First, a definition of the three levels of engagement.

- Low Engagement: This level of engagement is marked by an overall low intensity of focus on the exhibit by a majority of the visitors in the party in question. Causes for departure from the exhibit before the exhibit's completion included parental or teacher directives to leave, external distractions, or simply loss of interest.
- Medium Engagement: This level of engagement is characterized by a mix of physically observed signals indicating focus, or lack thereof, by the visitor interacting with the exhibit. Examples of such signals were whether the visitors were consistently looking at the exhibit screen or away from it, pointing to the screen, and predominance of chatter with other members of the party instead of focusing on the screen, etc. These signals indicated partial interest in the exhibit, but sufficient to not immediately abort the session and leave.
- High Engagement: This level of engagement is characterized by an overall high intensity of focus on the exhibit coupled at times with positive emotional responses from the visitor directed towards the exhibit (i.e. laughing, chattering, smiling, while looking at and/or pointing to the screen). We should note that not all highly engaged groups completed the exhibit, but their exhibit runs were consistent with the previously stated observations.

Engagement Levels by Constituencies

Table 7 summarizes the data correlating age group with level of engagement. In sample groups involving multiple visitors, the age group that fit the majority of visitors in the party who were present for most of the exhibit duration was selected. For parent-child groups, the dominant age group chosen was that of the child, as it was assumed that in most cases, the child was driving the exhibit experience.

Table 7 Age group and engagement level

	Low engagement	Med. engagement	High engagement	Total
Elementary School Age	6	8	8	22 (57.9%)
Middle School Age	0	0	2	2 (5.3%)
High School Age	1	2	2	5 (13.2%)
Young Adult (20's)	1	2	1	4 (10.5%)
Adult	2	3	0	5 (13.2%)
Column Total:	10 (26.3%)	15 (39.5%)	13 (34.2%)	38

Duration of Interaction for each Engagement Level

Table 8 below is a summary of the time on exhibit compared to engagement levels. All times are in minutes:seconds. We should note that although duration of interaction played a qualitative role in our individual classification of visitor experiences engagement, the actual measured duration was not directly used to classify an interaction experience. While there is some overlap, it appears from a qualitative examination that length of time of interaction correlates well with level of engagement. The average time of experience in the high engagement group closely approximates the duration of the exhibit to completion, indicating that most of them completed the exhibit.

Summary of Summative Evaluation

In conclusion, nearly 74% of the parties that activated and interacted with the exhibit demonstrated medium to high engagement with the exhibit. An attempt was made to survey some of these parties chosen randomly after their experience with the exhibit; however, most declined to be interviewed, citing time constraints. Therefore, this interview process was discontinued. The number of surveys completed (3) was much too small to be meaningful, and the data were discarded.

Summary

In summary, the assessment process in its formative mode was very helpful in improving the exhibit. In its summative mode, it reinforced that the choices made were effective. Nevertheless, there is room for further improvement in the exhibit as well as in its evaluation. The most salient point to come out of the summative tests is that the duration of the exhibit experience appears to be somewhat longer than optimal. This appears to be

Table 8 Summary of time breakdown for engagement levels (min:sec)

	Shortest time	Longest time	Shortest time (w/o outliers)	Longest time (w/o outliers)	Average time	Average time (no outliers)
Low engagement	00:10	01:48	00:15	01:40	00:54	00:52
Medium Engagement	01:00	05:23	01:20	04:45	02:47	02:43
High engagement	00:14	15.17	01:48	07:51	05:29	05:04
All levels of engagement	00:10	15.17	00:15	07.51	03:12	03:04

the case in spite of the fact that during formative testing, subjects consistently stated that its length was about right. While the question of appropriateness of duration was not asked directly in the summative tests, it became clear from the relatively high percentage of visitors who did not finish the exhibit that this is a problem. In retrospect, the difference in evaluation context - quiet back room with no pressure to move on to experience other exhibits vs. museum floor with other children and adults beckoning the visitor to move on - may have played a role in the different conclusions reached. We are currently looking at ways to shorten the duration of the exhibit without altering its primary message.

The three formative evaluations consisted exclusively of self-reporting surveys. While self-reporting is an effective means of assessment, it is subjective, especially so when the test subjects are children. Nevertheless, questions can be asked directly, and if the sample size is large enough, quantitative measurements such as with the Likert scale, can be meaningful. We sought to (informally) normalize the answers given by the underage test subjects by having a member of the research team administer the survey via an interview, often in the presence of a parent. While this researcher's interpretation of the subject's answer may have added another layer of subjectivity, it had the beneficial effect of effectively explaining to the children the questions being asked.

We should add that the formative evaluation program we undertook was more extensive than what is described in this paper. It included forming a panel of three outside reviewers who on a semi-annual basis reviewed documents being prepared by the research team and passed judgment on the effectiveness of our on-going work. These documents involved exhibit storyboards in the early stages of the work to videos of the prototype being executed (in lieu of having them actually experience the exhibit, which was not possible at the time). Their input and feedback was quite helpful in the early stages of the work – less so as the exhibit neared completion. Lastly, we video-recorded from a distance (30 ft away) all interactions with the exhibit by the formative assessment test subjects, as permitted under the approved IRB protocol. However, we found little use for those recordings.

For the summative assessment, we sought to use an alternative data collection approach, different from the self-reported data collection used during the formative evaluations. It was decided to use unobtrusive observation of the interaction of museum visitors with the exhibit. The visitors were unaware that they were being observed. For that reason, we did not electronically record anything of the interaction, nor electronically log their interaction with the exhibit in any way. After unsuccessfully asking a few early visitors to answer some questions, no further contact was made with any of them, thereby maintaining complete anonymity. The few data sets collected (3) were discarded. While this evaluation involved subjective observation, the presence of four observers who passed judgment on the level of engagement exhibited by the various parties of visitors provided a measure of objectivity.

In all, the combination of self-reporting during formative assessment and unobtrusive observation during the summative evaluation worked well and provided a good balance of data.

Conclusions and Comments

In retrospect, the project was quite ambitious given the limited funding obtained from the sponsor, as well as its constrained time frame. Administrative limits on the NSF CRPA

program capped the maximum funding available and the period of performance. This caused us to have to rely mostly on undergraduate research assistants, led by one doctoral student. To their credit, the students took this on as a challenge and produced an exhibit that remains highly popular with museum visitors nearly three years after initially placed there. Moreover, the exhibit software has suffered very few errors that have required new versions to be installed. In fact, the museum personnel report that the exhibit runs without trouble nearly all the time, only occasionally requiring a computer reboot. The biggest disappointment was that most visitors did not speak to the avatars, but rather choose to use the touch screen menus to make their decisions. This, we believe, attests to the modern digital culture of touch screens. In reality, however, a more modern/more effective ASR system would have to be purchased and installed before an effort is made to exhort visitors to speak to the avatars. The ASR system type selected was one of the first effects of the limited funding. Lastly, the museum was highly cooperative during testing and initial installation. Their design and construction of the exhibit kiosk was excellent.

Acknowledgements The research described in this article was funded by the US National Science Foundation under the Communicating Research to Public Audiences program, grant number DRL 1138325.

Appendix – Survey Questions used in Formative Assessment

1) You are in (check one): Middle school ____ High school: ____ College: ____ Post college ____

2) Your gender is (check one): Male ___ Female ___

Please answer the following questions on a 1 to 5 scale, where 1 = **positively No**; 2 = **mostly No**; 3 = **not sure either way/about right**; 4 = **mostly Yes**; 5 = **absolutely Yes**: and 0 = **not applicable or couldn't really tell**.

3) Was it too long? _____

4) Did you choose to interact with Eliza? Yes _____ No _____
   a. If yes, did you enjoy it? _____

5) Were you able to understand Turing's dialog? _____

6) Were you able to understand the dialog of the avatar you built? _____
   a. Which voice did you select? Anna ___ Mike ___ Crystal ___ Alain ___ Charles ___

7) Were you able to understand and follow the course of the exhibit? _____

8) Did you notice the background? (Yes/no) _____
   a. What do you think it was? _____

9) Did the exhibit flow smoothly? _____

10) Any changes you think would improve the exhibit?
    _____
    _____

11) Did the exhibit increase your interest in computer science/engineering? _____

12) Did you enjoy the exhibit? _____

(Note: The order of the questions was different in the surveys used for the three formative evaluations. We have adjusted the order above to make them consistent for ease of reporting. However, the wording of the questions was identical in all surveys.)

References

Becker, C., & Wachsmuth, I. (2004). Simulating the emotion dynamics of a multimodal conversational agent. *Proc. Tutorial and Research Workshop on Affective Dialogue Systems, ADS-04*.

Bell, P., Lewenstein, B., Shouse, A. W., & Feder, M. A. (Eds.). (2009). *Learning science in informal environments: People, places, and pursuits*. DC, National Academies Press: Washington.

Bickmore, T. W., & Picard, R. W. (2004). Towards Caring Machines. Computer Human Interaction. CHI'04 extended abstracts on Human factors in computing systems (pp. 1489–1492). ACM.

Bickmore, T., Pfeifer, L. & Schulman, D. (2011) Relational agents improve engagement and learning in science museum visitors. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. 6895: 55–67. LNCS Springer: Heidelberg

Brézillon, P. (2003). Representation of procedures and practices in contextual graphs. *The Knowledge Engineering Review., 18*(2), 147–174.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., & Yan, H. (2000). Human conversation as a system framework: designing embodied conversational agents. *Embodied conversational agents,* (pp. 29–63).

DeMara, R. F., & Moldovan, D. I. (1991). Performance Indices for Parallel Marker-Propagation. Proceedings of the 1991 International Conference on Parallel Processing (ICPP-91). pp. 658–659.

DiSalvo, D. (2012). How Alan Turing Helped Win WWII And Was Thanked With Criminal Prosecution For Being Gay. Forbes Magazine. http://www.forbes.com/sites/daviddisalvo/2012/05/27/how-alan-turing-helped-win-wwii-and-was-thanked-with-criminal-prosecution-for-being-gay/#1d3793812826. Accessed 27 May.

Duch, W., Szymański, J., & Sarnatowicz, T. (2006). Towards avatars with artificial minds: Role of semantic memory. *Journal of Ubiquitous Computing and Intelligence, 1*.

Facegen (2012). Facegen Modeller: 3D face generator. Retrieved from Facegen.http://www.facegen.com/modeller.htm. Accessed 24 Sept 2012.

Falk, J. (2011). Webinar, April 19, 2011. Sponsored by Education Week Magazine.

Foner, L. (1993). What's An Agent, Anyway? A Sociological Case Study. Vol. 1. Agents Memo 93.

Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellener, B., Simmons, R., Snipes, K., Schultz, A. C., Wang, J. (2005). Designing robots for long-term social interaction. IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1338–1343.

Gonzalez, A. J., DeMara, R. F., Hung, V. C., Leon-Barth, C., Elvir, M., Hollister, J. R., Soros, L., Kobosko, S., Leigh, J., Johnson, A., Jones, S., Carlson, G., Lee, J., Renambot, L., & Brown, M. (2013). Passing an enhanced Turing test – Interacting with lifelike computer representations of specific individuals. *Journal of Intelligent Systems., 22*(4), 365–415.

Hayes-Roth, B., & Doyle, P. (1998). Animate characters. *Autonomous Agents and Multi-Agent Systems, 1*, 195–230.

Hollister, J. R., Parker, S. L., Gonzalez, A. J., & DeMara, R. F. (October 2013). *2013*. A Context Based Approach Designed to Educate Youth in Computing. Context Conference. Annecy, France: An Extended Turing Test.

Hung, V. C., & Gonzalez, A. J. (2013). Context-centric speech-based human-computer interaction. *International Journal of Intelligent Systems, 28*(10), 1010–1037.

Johnson, W. L., Rickel, J., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education., 11*, 47–78.

Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsela, S., Piepol, D. (2007). Building Interactive Virtual Humans for Training Environments. I/ITSEC'07.

Kopp, S., Gesellensetter, L., Krämer, N. C., & Wachsmuth, I. (2005). A conversational agent as museum guide – Design and evaluation of a real-world application. *Intelligent Virtual Agents. Lecture Notes in Computer Science., 3661*, 329–343.

Lane, H. C., Cahill, C., Foutz, S., Auerbach, D., Noren, D., Lussenhop, C., & Swartout, W. (2013). The effects of a pedagogical agent for informal science education on learner behaviors and self-efficacy. *Proceedings of the AIED* 2013. *LNAI, 7926*, 309–318.

Lee, C., Sidner, C., Kidd, C. (2005). Engagement during dialogues with robots. AAAI Spring Symposia.

Mateas, M. (1997). An Oz-centric review of interactive Drama and believable agents. In J. G. Ccarbonell & J. Siekmann (Eds.), Artificial intelligence today: Recent trends and developments. Berlin: Springer Berlin/ Heidelberg.

National Research Council. (2005). Rising above the gathering storm: Energizing and employing America for a brighter economic future. In *Committee on prospering in the global economy of the 21st century; committee on science, engineering, and public policy; division on policy and global affairs*. The National Academies, Norman Augustine: Chair.

Shiomi, M., Kanda, T., Ishiguro, H., Hagita, N. (2006). Interactive humanoid robots for a science museum. *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction,* 305–312.

Sorge, C. (2007). What happens? Relationship of age and gender with science attitudes from elementary to middle school. *Science Educator, 16*(2).

Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., Lane, C., Morie, J., Aggarwal, P., Liewer, M., Chiang, J.-Y., Gerten, J., Chu, S., & White, K. (2010). Ada and Grace: Toward realistic and engaging virtual museum guides. *Proceedings of the IVA 2010 Conference*. *LNCS (LNAI), 6356*, 286–300.

Traum, D., & Rickel, J. (2002). Embodied agents for multi-party dialogue in immersive virtual worlds. Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 2.

Traum, D., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., Leuski, A., Noren, D. and Swartout, W. (2012) Ada and Grace: Direct interaction with museum visitors. *International Conference on Intelligent Virtual Agents,* 245–251.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

Weizenbaum, J. (1966). ELIZA-a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45.