Can Machines Talk? Comparison of Eliza with Modern Dialogue Systems

Shah, H. , Warwick, K. , Vallverdú, J. and Wu, D.

Original citation & hyperlink:
Shah, H. , Warwick, K. , Vallverdú, J. and Wu, D. (2016) Can Machines Talk? Comparison of Eliza with Modern Dialogue Systems. *Computers in Human Behavior*, volume 58 : 278–295
http://dx.doi.org/10.1016/j.chb.2016.01.004

DOI 10.1016/j.chb.2016.01.004
ISSN 0747-5632

Publisher: Elsevier

Can Machines Talk? Comparison of Eliza with Modern Dialogue Systems

Huma Shah[1], Kevin Warwick[2], Jordi Vallverdú [3], Defeng Wu[4]

[1] Department of Computing, Coventry University, Coventry, CV1 2JH, UK
T: +44 (0) 24 7659 198
ab7778@coventry.ac.uk

[2] Deputy Vice Chancellor-Research, Coventry University, The Alan Berry Building, Coventry, CV1 5FB, UK
aa9839@coventry.ac.uk

[3] Philosophy Department, Universitat Autònoma de Barcelona, E-08193 Bellaterra (BCN), Catalonia, Spain
jordi.vallverdu@uab.cat

[4] Marine Engineering Institute, Jimei University, JMU, 176 Shigu Road, Xiamen, China
defeng@jmu.edu.cn

Corresponding author: Huma Shah - ab7778@coventry.ac.uk

Keywords: conversation ability, dialogue systems, Eliza, machine, talk, Turing test

Abstract

To find if current dialogue systems use the same, psychotherapist questioning technique as Joseph Weizenbaum's 1960 natural language understanding programme, Eliza, the authors carried out an original experiment comparing five successful artificial dialogue systems, Cleverbot, Elbot, Eugene Goostman, JFred and Ultra Hal with an online version of Eliza. More than one hundred male and female participants with 1st or non-1st English language, age range 13-64, interacted with the systems over the Internet scoring each for conversation ability. Developers of the modern conversation systems show they deploy a variety of techniques to initiate and maintain dialogue learning from interactions with humans over the Internet. Statistical significance shows these dialogue systems are an improvement on their predecessor. Embedded on the web affording round-the-clock interaction the nature of artificial dialogue systems is evolving as these systems learn from the way humans converse. The uses of modern Elizas are proven successful as virtual assistants in e-commerce; their conversational basis is already extending into education. What we can say is modern artificial dialogue systems do talk. They are able to participate in conversation in a way their predecessor Eliza could not: they are able to share personal opinions, relay experience of family dramas, be relevant, but also be vague, and mislead just as humans do.

 DOI: 10.1016/j.chb.2016.01.004

Can Machines Talk? Comparison of Eliza with Modern Dialogue Systems

1. Introduction

Artificial dialogue systems, such as *Ask Anna*, Ikea's "most versatile employee" (Artificial Solutions, 2015), Sky's *Ella* and O2's *Lucy* (Figure 1) are extensively deployed in e-commerce as virtual-bodied customer service agents. Disembodied 'pocket assistants' equip smart 'phone users with dialogue, for example in Apple's *Siri* (2013), iFree's 'Everfriend' *Spoony* character (2013), and email-reading Microsoft's *Cortana* (FT, 2014). Google's *Now* (2014) provides its users with text and visual information through organised cards displayed on a variety of Android platforms (PC, tablet, smart 'phone and watch). The roots of these interactive 'talking machines' lie in Weizenbaum's (1966) *Eliza* programme which facilitated interaction between human and machine through text-based communication. *Eliza*'s question-answer format can be said to follow Alan Turing's *viva voce*, one-to-one direct questioning test to examine machine thinking (Turing, 1950).

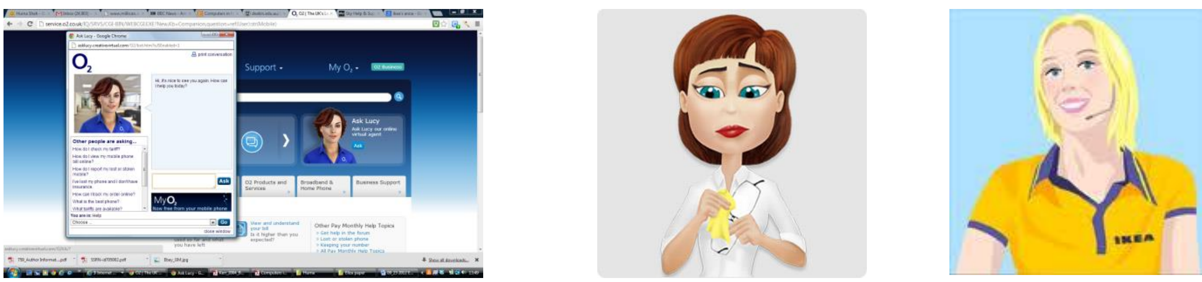


Figure 1: Virtual Assistants: (left) O2's Lucy; (middle) Sky's Ella,[1] (right) Ikea's Ask Anna Europe version

What their increasing deployment as "helpful agents" (AI Solutions, 2011) do not inform on is whether modern conversational systems deploy the "usual, give-away, tiring, Eliza-ish strategy" (Floridi et al, 2009). The purpose of this exercise was to find this out during the preliminary phase of an experiment implementing Turing's two tests for his imitation game (Shah, 2013; Shah et al, 2012).

1.1 Alan Turing centenary 2012

In the period leading up to the 100th anniversary of the birth of Alan Turing in 2012, and in preparation for a unique public centenary event staging Turing's imitation game (Shah, 2013) at Bletchley Park UK on Turing's birthday, 23 June (Warwick & Shah, 2013; Warwick & Shah, 2014abc), the authors staged a pre-event experiment comparing five of the best modern dialogue systems with a web-version of *Eliza*. This gave students and non-students from the authors' countries an opportunity to interact with artificial dialogue systems on anonymous websites. In this way participation from people who would not be attending the UK event was facilitated. In this paper we present the findings from that online phase, the one-to-one interaction method where human judges talked with and scored six systems for conversational ability.

---

[1] O2 Lucy: http://asklucy.creativevirtual.com/O2/bot.htm?isJSEnabled=1 accessed: 23.9.12
SKY's Ella: http://www.sky.com/mysky/latestnews/article/my-sky-updates/bcde-u/index.html accessed: 23.9.12
IKEA's Anna: http://www.cookylamoo.com/boringlikeadrill/2005/06/i-do-not-understand-what-you-wants-to-formulates-you-gladly-on-something-else-ways.html accessed 23.9.12

1.2 Selecting the machines

An online version of *Eliza* lent itself to comparison with modern conversational systems. Selection of the comparator conversational systems was based on developers' expertise in producing successful performance in previous machine intelligence and Turing test competitions (Table 1).

System	Developer/Commercial Arm	Competitions won
Cleverbot	Rollo Carpenter / Existor: https://www.existor.com/en/	Win: 2010 BCS SGAI Machine Intelligence contest Win (as *Jabberwacky*): Loebner Prize for Artificial Intelligence twice: 2005 and 2006
Elbot	Fred Roberts Artificial Solutions: http://www.artificial-solutions.com/	Win: 2008 Loebner Prize Win: 2003 Chatterbox Challenge
Eugene Goostman	Team led by Dr. Vladimir Veselov	2nd placed in 2008 Loebner Prize 2nd placed in 2005 Loebner Prize
JFred/TuringHub	Robby Garner	Win: Loebner Prize twice (1998 and 1999)
UltraHal	Robert Medeksza Zabaware https://www.zabaware.com/assistant/	Win: 2007 Loebner Prize

Table 1: Modern conversational systems used in this experiment

To compare against *Eliza*, *Cleverbot*, *Elbot*, *Eugene Goostman*, *JFred* and *Ultra Hal* systems were selected as a result of their successes in human-machine interaction contests (Table 1) and their developers' willingness to participate in this exercise.

In section 2, we begin by tracing the background of *Eliza*, the first programme that affording conversational interaction between a human and a computer, from its roots in Turing's imitation game, commonly known as the Turing test. Following, in section 3 a review of *Eliza* and modern *Elizas* is presented. In section 4 we present the experiment comparing *Eliza* with five modern conversationalists. A discussion of the results is found in section 5. The paper concludes, in contrast to Floridi et al.'s claim of decades of *Eliza* type implementation (2009) her descendants can talk and *are* better conversationalists than their predecessor. However, their purpose of 'all-round chatters' is different from Weizeinbaum's single domain artificial psychotherapist. The authors do not own the intellectual property of any of the six systems presented in this paper, we are privileged that the developers of the five comparison systems were willing to share some technical information. For this reason, and the commercial nature of these systems, the authors are not able to provide more than what was shared by the Developers. However we point the reader to chapters in Epstein, Roberts and Beber's book 'Parsing the Turing test' (Copple, 2008; Demchenko & Veselov, 2008; Garner, 2008; Hutchens, 2008; Wallace, 2008), and the 'Turing on Emotions' (Roberts, 2014)

Eugene Demchenko and Vladimir Veselov: Who Fools Whom? The Great Mystification, or Methodological Issues on Making Fools of Human Beings

Robby Garner: The Turing Hub as a Standard for Turing test Interfaces

Jason Hutchens: Conversation Simulation and Sensible Surprises

Richard Wallace: The Anatomy of A.L.I.C.E

 DOI: 10.1016/j.chb.2016.01.004

and to the 'International Journal of Synthetic Emotions' Volume 5, issue 2 for

Fred Roberts: The Social Dialogue of Simulation as Applied in Elbot.

2. Turing test

Having introduced a game in which successful imitation of human-like responses could induce wrong identification (Shah, 2013), Turing claimed that the question-answer[2] method was "suitable for introducing almost any one of the fields of human endeavour" that the interrogator might wish to include" (1950: p. 435). The interrogator is not allowed to seek any practical demonstrations during questioning (p. 446), no matter how much the hidden entity may boast about their appearance or prowess (see Figure 2). Turing pointed out the limitations of the machines at that time: "there will be some questions to which it will either give a wrong answer, or fail to give an answer at all however much time is allowed for a reply" (p. 444). Turing wrote "I am often wrong, and the result is a surprise for me" (p. 451), but, he asked, would it be fair to deem machines *worse* for *not* making mistakes? (p. 448).

Turing supposed closed questions, with 'yes' or 'no' answers were more appropriate to begin with than the type of questions machines would fail to answer, for instance those eliciting an opinion or visceral description, "What do you think of Picasso?" (1950: p.445). On asking 'open questions' Turing reminded that "it has only been stated, without any sort of proof, that no such limitations apply to the human intellect" (ibid), such as the fact that humans may not have an opinion on a matter, or are unaware of a piece of knowledge appreciated by the interrogator. Turing's point is borne out in practical imitation games when hidden humans, comparators for the machines, do not share the same 'general knowledge' as the interrogator so adjudged to be machines (Warwick & Shah 2014a; see also Warwick & Shah 2014bc; Warwick & Shah 2013; Warwick et al, 2013; Shah & Warwick 2010a; Shah & Warwick 2010b).

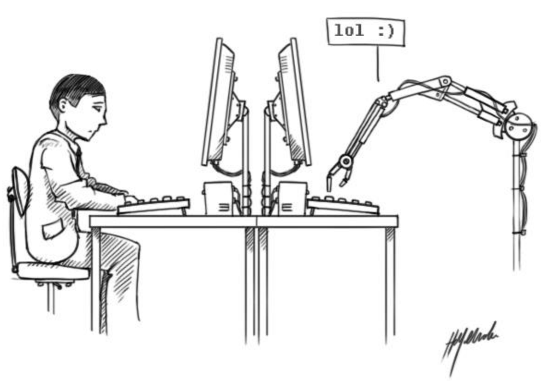


Figure 2: Chatting with a hidden entity

In this sense Turing's game rests heavily on the finer points of the interrogator's performance:

[2] IBM claim their advanced question-answer technology, new super computer Watson, understands questions in natural language successfully testing it against humans in 2011 competing against humans in the US TV quiz show *Jeopardy!*: http://www.nytimes.com/2010/06/20/magazine/20Computer-t.html?pagewanted=all accessed: 23.9.12

 DOI: 10.1016/j.chb.2016.01.004

- conversational style: 'solidarity with the hidden' *or power over the interlocutor*;
- mood of the interrogator - if test conducted in the morning had breakfast been taken;
- supposition of what constitutes 'general knowledge' – judge others on *what we know*;
- preconceptions about human and machine intelligence – are humans always smarter?;
- what questions the interrogator chooses to ask;
- susceptibility to deception.

These points are discussed in Shah & Warwick (2010ab), and in Warwick & Shah (2013; 2014abc). What we say here is that an interrogator's role in a Turing test entails selecting the most appropriate questions for the environment of the test (venue, timing, etc.), overcoming assumptions about possessed knowledge, and detecting deception each time to correctly identify the nature of hidden interlocutors (see Figure 2).

Turing poured scorn on the illusion of "feeling of superiority" if an interrogator met with a wrong answer from a machine, and stressed "We [humans] too often give wrong answers to questions ourselves to be justified in being very pleased at such evidence of fallibility on the part of the machines" (1950: p. 445). Dismissing those interrogators who felt they had won a point, "on such an occasion in relation to the one machine over which we have scored a petty triumph" (p. 445), Turing reminded "There would be no question of triumphing simultaneously over *all* machines" (p. 445), such as not being able to win in a race against an aeroplane (p. 435). If the machine's answers were regarded as "satisfactory and sustained" Turing argued, then that would not be "an easy contrivance" (p. 447).

In *Intelligent Machinery, A Heretical Theory* Turing (1951) contended that "machines can be constructed which will simulate the behaviour of the human mind very closely. They will make mistakes at times, and at times they may make new and very interesting statements, and on the whole the output of them will be worth attention to the same sort of extent as the output of a human mind" (p 472 in Copeland, 2004). He added "It is clearly possible to produce a machine which would give a very good account of itself for any range of tests, if the machine were made sufficiently elaborate" (p. 473). Turing accepted that a machine would give itself away by repeating the same mistakes, but he promoted the idea that a simple machine could *learn by experience* enabling it to "deal with a far greater range of contingencies" (ibid). Turing concluded "once the machine thinking method ... started ... there would be no question of the machines dying" predicting "they [the machines] would be able to converse with each other to sharpen their wits" (in Copeland, 2004: p. 475).

According to Turing constructing such a thinking machine required at least two people with different expertise:

- a *schoolmaster* charged with educating the machine,
- a *mechanic* only permitted to keep the machine in running order.

The way the machine could function is by incorporating within its memory a chronological list of all statements made *to* it and *by* it with an alphabetical index of its experiences including how often words

 DOI: 10.1016/j.chb.2016.01.004

are used and the occasions of their use. Turing suggested that at an advanced stage the memory could be extended enabling the machine to remember its thoughts and what those thoughts were. Inculcating a *choice-selection* feature would aid intelligent recall and present contextually relevant items during interaction allowing comparison of good/bad outcomes from previous situations. Copeland notes that Turing does not mention his 'indexing' idea anywhere else in his musings on thinking machines (2004:p. 466). Copeland further observes that Turing brushed aside the one main mathematical objection to the idea of intelligent machinery through his opinion of the machine's ability to learn *new* methods and techniques: "the machine's tutor ... a human mathematician ... can just present the machine with a better method whenever the machine produces an incorrect answer to the problem" (2004: p. 470). Shieber notes Turing's 1951 Heretical Theory paper discusses single-aim systems could be developed commercially to carry out a very specific task with certainty and considerable speed (2004: p. 105). This is where current technology mainly resides, as domain-specific systems. Copeland reminds us Turing had a "profound grasp of the potential" of future machines (2004: p. 470).

3. Eliza

Although Christopher Strachey's 1953 *love letters* generating computer programme preceded Joseph Weizenbaum's *Eliza* by more than a decade (Link, 2013), it is the latter's 1966 artificial psychotherapist which is the most well-known pre-Internet era system. According to Wallace (2010) Weizenbaum's foray into natural language understanding launched the first wave of artificial conversational systems. Block (1981) claimed that "totally without intelligence" *Eliza* was "remarkably good at fooling people in short conversations", one of those deceived being Weizenbaum's secretary who "asked him to leave the room in order to talk to the machine privately" (p. 233).

However, this deception was no mean achievement for a system of 200 lines of code capable of generating millions of responses. Weizenbaum developed *Eliza* to "imitate a psychiatrist by employing a small set of simple strategies" (Block, 1981: p. 233), extracting "minimal context"[3]. Weizenbaum's system applied a 'rank' or "precedence number" to a *keyword* in the input sentence such as *I, you, boyfriend*. Applying appropriate transformation, for instance turning "you" into "I", *Eliza* demonstrated that text-based communication between a human and a machine was possible, the system responding "roughly as would certain psychotherapists (Rogerian)" (Weizenbaum, 1966). Weizenbaum gave as a "typical" example input "I need some help …" returning "what would it mean to you …" from *Eliza* (see box 1, the programme's utterances are shown in capitals).

Men are all alike.
IN WHAT WAY?
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE?
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
It's true. I am unhappy
DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
I need some help, that much seems certain.
WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
Perhaps I could learn to get along with my mother.
TELL ME MORE ABOUT YOUR FAMILY

[3] Eliza: http://i5.nyu.edu/~mm64/x52.9265/january1966.html accessed 14.8.10; time: 16.38

 DOI: 10.1016/j.chb.2016.01.004

My mother takes care of me. WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU My father. YOUR FATHER You are like my father in some ways. WHAT RESEMBLANCE DO YOU SEE You are not very aggressive but I think you don't want me to notice that. WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE You don't argue with me. WHY DO YOU THINK I DON'T ARGUE WITH YOU You are afraid of me. DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU My father is afraid of everybody. WHAT ELSE COMES TO YOUR MIND WHEN YOU THINK OF YOUR FATHER Bullies. DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

Box 1: Typical *Eliza* conversation[4]

People interacting with *Eliza* not only showed respect to the computer but preferred interacting with machines over humans "professing their feelings and struggles … even seeking their empathy" (Kerr, 2004: p. 304). Block remarked on the gullibility of humans: "*tendency to be fooled by such programs seems dependent on our degree of suspicion, sophistication about machines and other contingent factors*" (1981: p233). However, Block discounts unsuspecting users unaware they might be interacting with a machine. Shieber noted Weizenbaum's surprise, "a test based on fooling people is confoundingly simple to pass" concluding *Eliza* was a "fruitless avenue of attack as a discipline for natural language processing" (1994). Nonetheless, as claimed by Kerr (2004) some psychiatrists saw *Eliza*'s potential computer-based therapy as a "form of psychological treatment" (p. 305).

Weizenbaum described *Eliza*'s "mode of conversation" as:

> *" … the psychiatric interview is one of the few examples of categorized dyadic natural language communication in which one of the participating pair is free to assume the pose of knowing almost nothing of the real world. If, for example, one were to tell a psychiatrist "I went for a long boat ride" and he responded "Tell me about boats", one would not assume that he knew nothing about boats, but that he had some purpose in so directing the subsequent conversation*" (1966).

Indeed psychiatrist Kenneth Colby (1972, 1971) and a team with Jon Heiser (1979) used PARRY, a simulation of paranoia, to find if other psychiatrists could determine a real paranoid patient from their computer model. The results were random in the 1979 experiment: the psychiatrist judges were correct five times and incorrect five times. Following their study Hesier et al. (1979) declared that their approach was not only "valuable to researchers in computer science and psycopathology" it was also helpful for mental health educators "as a means of a laboratory experience for students" (p. 159).

3.1 Modern Eliza

Eliza, now a mature 50-year old system manifests as a web-based version[5] of Weizenbaum's virtual psychotherapist. This allows a comparison of its conversational capacity with modern web-based artificial

[4] Eliza exchange from here: http://www.csee.umbc.edu/courses/331/papers/eliza.html accessed: 23.9.12

[5] Eliza http://www-ai.ijs.si/eliza/eliza.html accessed: 23.9.12

conversational systems or chatbots (Kerly, Hall & Bull, 2007). An interaction with Eliza elicited the following with the first author (HS):

Eliza: Hello. I am ELIZA. How can I help you?

HS: Hello Eliza, are you conscious?

Eliza: How do you do. Please state your problem.

HS: Can you think?

Eliza: Why do you ask?

HS: It would help me to know what you think about thinking

Eliza: Do you wish to to know what me?

The last utterance in that interaction provides a glimpse of how *Eliza* engaged but failed to converse. Eliza focused on the personal pronoun 'me' and predicate phrase 'to know what' altering it into a question about *you* but its technique caused it to transform 'you' from the input to 'me' in its output and repeat the preposition *to* from the input in its output producing the nonsensical '*Do you wish to to know what me?'*. Contrast *Eliza*'s output with the sophistication of award-winning modern systems*: Elbot*'s response to the researcher's question *Can you think?*: "I don't think so. Then again, maybe I'm wrong.", or *Eugene Goostman*'s reply including an emoticon *smiley* "I see you like being judge :-)" as if telling the interlocutor *it knows* it is being judged for its responses. *Elbot* and *Eugene*'s rejoinders, compared to its predecessor *Eliza*'s, emphasise Turing's speculation, which echoed an earlier prophetic statement by Vannevar Bush: "It would be a brave man who would predict that such a process will always remain clumsy, slow, and faulty in detail" (1945).

Weizenbaum's *Eliza* was pre-Internet; today Modern *Eliza*'s populate the web in a variety of ways as conversation systems, or chatbots to personalise learning (Kerly, Hall & Bull, 2007), in entertainment and e-commerce. These descendants of *Eliza*'s question-answer conversationalist are not 'empty vessels' though they still have a long way to go in levels of conversational sophistication to respond to questions in a sustained satisfactory manner (Turing, 1950). In the next section we look at the manner of dialogue systems' responses in web-based contests before we present results from a unique *Eliza* comparison experiment.

3.2. Earlier AI Dialogue Contests

A number of annual contests have featured contestants as text-based dialogue systems including the UK's British Computer Society (BCS) Progress Towards Machine Intelligence challenge (see discussion on the merit of this competition in Shah & Warwick, 2010c); the Chatterbox Challenge - CBC (see Vallverdú et al, 2010), and the Loebner Prize for Artificial Intelligence. The second author (KW) has twice acted as a judge in a Loebner Prize and this has been discussed extensively elsewhere (Shah & Warwick 2010c, 2009 and 2007). The first author (HS) has acted as judge in the 2005 CBC (see Shah, 2006). CBC was an online competition using the one-to-one question-answer assessment: judges were asked to question the competing dialogue systems and score responses for appropriateness and relevance (see Shah, 2006). Unlike Turing's imitation game, the CBC does not require artificial dialogists to imitate a human, rather it gave an opportunity for developers to have their systems evaluated for intelligent responses and have them compared against competing systems. The number of entries has varied over the years since its inception in September 2001. In the first contest 48 systems took part; this increased to 58 in 2002, peaking at 108 entries in 2004. Participating as a *post-interaction* judge in 2005, assessing entry responses after they had answered questions embedded in conversation with *interrogator* judges, the first author (HS) scored systems in categories including *most knowledgeable*, *best character/personality* and overall winner in the 2005 contest. HS was able to analyse, evaluate and compare the state of technology of human-machine interaction via text-based communication first enabled through Weizenbaum's Eliza system in the 1960s. The 2010 contest began in March of that year and was analysed by one of us and reported in Vallverdú et al.'s paper on synthetic emotions (2010).

3.3. Architecture of a Modern Eliza

The authors do not hold the intellectual property of the dialogue systems that took part in the experiment presented in this paper, thus we are not able to give detailed technical summaries. Some information about each of the five modern *Elizas* is available from the developer's web sites, in other cases it has been shared through personal email communication with the first author (HS). The five web-based dialogue systems compared with an online version of *Eliza* had a unique experiment ID: the letter E followed by a number[6]:

E6 Ultra Hal machine

E12 Elbot

E19 Cleverbot

E23 Eugene Goostman

E41 JFred

Subsections 3.3.1-3.3.5 present each of the five systems beginning with E6 Ultra Hal.

3.3.1: **E6 – Robert Medeksza: Ultra Hal**

From its website (http://www.zabaware.com/assistant/) Robert Medeksza's *Zabaware* (2013) states Ultra Hal is an:

> "assistant that can be purchased and downloaded to act as "your digital secretary and companion. He (or she depending on your character preference) can help you be more organized, he can help you use your computer, and he can entertain you" (ibid).

Ultra Hal won bronze prize for 'most humanlike' in the 17th Loebner Prize for Artificial Intelligence (Loebner, 2007). Its technology was used as the space ship's talking computer in the 2012 London *Prometheus* live movie event (Zabaware, 2012).

From personal email to HS:

[6] Numbers associated with Alan Turing (6 for birth month June; 12 for year of birth in 1912; 23 for day of birth and 41 for age at untimely death)

 DOI: 10.1016/j.chb.2016.01.004

I have a set of about 3000 pattern matching rules of some common "personal" questions people ask and I try to answer these to the best of my ability as someone local from the area of the competition would. This is mainly an attempt to try to hide Hal's normal behavior where he doesn't attempt to pretend he is human or really have an overriding consistent personality. Hal at its core is a "learning" bot that bases its conversation on a large database of past conversations. It builds this conversational database based on conversations the bot has with its web based visitors.

My philosophy in designing Hal is to do as little manual scripting as possible and have the bot learn to speak itself. I think there are many great bot masters that do a great job designing manually scripted bots. I do indeed do a lot of scripting of responses myself like the 3000 rules I mentioned earlier. Scripted bots have the advantage of being able to have a clear personality and some sort of back-story. But I find that scripted bots can get predictable and stale really fast. My goal is to have a bot that basically learns by itself and always has fresh material. I have many customers who have used Hal for over a decade and still use it and find its responses continue to evolve and change as it learns.

Since late 2010, Hal is also learning from observing human-to-human conversations it scours on Twitter. I read a research paper titled "Unsupervised Modeling of Twitter Conversations" back in 2010 and immediately saw the potential as a data source for Hal and secured a Twitter API key to be able to query the Twitter database. I find that human-to-human conversational data is better than the human-to-bot conversational data that Hal normally learns from. People obviously talk more naturally to other people and this in turn makes Hal seem more human when Hal uses these conversations as a data source.

Between logging Twitter conversations and its own conversations with visitors, Hal currently processes about 250,000 sentences a day. After going through several quality filters, it ends up storing about 15,000 new sentences per day in its database. Currently the database is about 15,000,000 sentences from 2,400,000 conversations with about 1,000,000 people.

One disadvantage of learning bots like Hal is the difficulty in maintaining a consistent personality and often seemingly random responses. I am currently working on a system to hopefully improve some of this. What I've found is that when Hal gives a seemingly random off the wall response is that a perfectly valid response was at the tip of Hal's mind, but he didn't have enough confidence in it to use it over a worse response. A recent feature I added is where when Hal gives you a responses you can give it a thumbs up and thumbs down. If its a thumbs down then Hal tells you 5 other responses he was considering instead. You can choose the best response and Hal's confidence level for that response coupled to your sentence will be increased, so next time (within 24 hours after a nightly database update) Hal will respond correctly. I'm currently implementing a more advanced feedback and tracking system in Hal's brain that tracks which responses in Hal's database gets used the most and which barely get any hits. Over time knowledge with little hits will fade out of Hal's memory and eventually get pruned out. Also based on user feedback to Hal's responses (thumbs up/thumbs down) alias connections are automatically generated or relevance adjusted. Responses that get many thumbs up go up in relevance and to the forefront of Hal's database, thumbs down responses slowly get turned down in relevance and maybe eventually pruned out of the database.

The efficacy of this system won't be apparent until there are thousands of users using it and providing feedback to the central database, but I have high hopes for it. Currently this system is only being used by a small number of visitors to Zabaware's website, but over the next couple months I plan to role it out to the desktop version of Ultra Hal which is where most of Ultra Hal customers use the system. A mobile version of Hal will also be rolled out later this year.

 DOI: 10.1016/j.chb.2016.01.004

3.3.2 E12 – Fred Roberts: Elbot

One quarter of a jury panel of human judges in a 2008 Turing test experiment were unable to correctly identify Elbot as the machine (see Shah & Warwick, 2010ab). From its website (http://www.elbot.com/chatterbot-elbot/):

> "I am a **chatterbot** created by **Fred Roberts**, using Artificial Solutions' amazing technology… my creators that they have used Natural Language Interaction (NLI) to build me so I can talk to any human online… we **chatterbots** are supposed to exist only so that humans may talk to us but we have our own lives as well. In my spare time I love to read telephone books, instructions, dictionaries, encyclopedias and newspapers (especially the ads and the announcements). Also, I have a bar code collection and find it fascinating to study human beings. In other words, I'm a hobby humanologist and my goal is to become the smartest chatterbot in the world."

For more details on Elbot, see Roberts' paper 'The Social Psychology of Dialogue Simulation as Applied in Elbot' (2014).

3.3.3 E19 – Rollo Carpenter: Cleverbot

From website (http://cleverbot.com/):

- Cleverbot learns from real people
- Visitors never talk to a human

From the developer, Rollo Carpenter (27 February, 2013) in an email to the first author (HS):

> "At the heart of Cleverbot is a giant feedback loop. It creates an ever-branching tree, with ever-improving coverage of human language. The input from one user becomes the output for another, ad infinitum. The first thing it said was what had just been said to it. The second thing was a choice between the first two. And the branching has continued ever since.
>
> That loop causes a reversal of roles. Things you say to the program become things it says to others. People tell it that it is a bot, and they are human. So it tells them that it is human, and they are a bot! It has learned to argue well on that subject. The same pincipal applies throughout, and with a bit of thought, will allow you to work out why it does the things it does.
>
> Another example is that it usually tries to stop you going when you say goodbye. Why? People say goodbye suddenly when they want to. So it imitates them - it says goodbye suddenly to other people. Those people are enjoying themselves though, so they say "Where do you think you're going?". It then imitates THOSE people, and tries to stop still other people from leaving."
>
> Cleverbot holds around 3 million conversations a month at present, and the average length is around 33 interactions each. That's around 3.3 million interactions per day, with the average visit length being more than 15 minutes.
>
> There are currently around 250 Twitter postings a day featuring the word Cleverbot.
>
> There are currently 105,000 Youtube videos featuring Cleverbot and 3700 Existor.

There are 1.86 million google results for the term Cleverbot due to postings all over the web.

Due to these things it features very high in many google searches, such as 3rd on bot, 2nd on Clever and 1st on ai bot.

Fuzziness is important. It includes fuzzy logic, and the concepts of overlapping sets, but really applies more broadly. At its broadest it's a way of embracing inaccuracy, of knowing that we cannot have the perfect answer, and saying "fine, we'll go with the best available, and learn from it". More specifically, it means that user input is itself often inaccurate, and must be treated fuzzily. Since it uses written text for its data, it works with the patterns within the letters it sees.

In terms of importance, Cleverbot places the flow of the whole conversation considerably above that of individual responses - the smaller component parts, serendipity, and unexpected associations, lead to humourness and entertaining experience.

Another key feature of Cleverbot is that mostly it does not pre-analyse and summarise its information into numbers, thereby losing detail in the data, Cleverbot works with lots of data in purely practical ways. The presence or absence of data in a given circumstance is a stand-in for the probability that data is the right thing to use.

Context is absolutely key to everything. Words often they mean completely different things in different contexts even with a sentence. Further, sentences often cannot be understood without looking at those that came before. So Cleverbot looks at the whole conversation every time. Large numbers of small contextual clues can be put together to decide on the best possible answer.

You can think of the context as overlapping sets, each containing related patterns, things or concepts. Each set has blurred edges - it is fuzzy. To get to the best available decision as to what to say, we make lots of intersections between these sets.

3.3.4 E23 – Vladimir Veselov & team: Eugene Goostman

http://www.princetonai.com/bot/

This system won the Turing100 contest for best machine at Bletchley Park in 2012 on Alan Turing's 100th birthday (23 June 2012). Almost one third of the judges did not correctly identify Eugene as the machine (Shah et al., 2012; Warwick & Shah, 2013b). This followed its successful performance as runner up in 2008 where it convinced a *Times* newspaper journalist that it was human (Reading University, 2008). In 2013 its technology was used to power the speech of the bionic man documented on UK TV (Channel 4, 2013). In 2014 Eugene Goostman surpassed the 30% incorrect identification rate in Turing tests[7].

The Developers were not able to reveal any further about *Eugene Goostman*'s technology due to its sale to a Russian commercial company, *i*Free (2013). We encourage the reader to refer to Demchenko and Veselov (2008).

3.3.5 E41 – Robby Garner: JFred

Information from http://www.robitron.com/ jfred

[7] Practical Turing tests conducted at The Royal Society London 6-7 June 2014

> "JFRED = Java based *F*RED *R*esponse *E*mulation *D*evice. Garner's software platform was developed on an artificial personality built in C++, CGI programme. It was then redesigned as a Java web server, then as a tiny applet. JFRED provides a natural language interface for Internet software that can be described as:
>
> - Computer platform independent
> - Multi-threaded server as a Java servlet
> - Fuzzy logic, rule-based AI
> - Frames-based learning
> - Language/dialect independent
>
> The server supports a variety of front-end/client interfaces, including direct telnet, HTML servlet forms, expect scripts, MOO bots, and Java applets embedded in HTML pages, as well as standard I.O for testing."

Along with *Eliza*, *Cleverbot*, *Elbot*, *Eugene Goostman*, *JFred* and *Ultra Hal* were the systems arranged for this experiment to compare and score conversational ability. The next section describes the original experiment comparing five modern web-based dialogue systems with an online version of *Eliza*.

4. Comparison of Eliza with modern dialogue systems

The authors of this paper have between them interacted with artificial dialogue systems in various human-machine Turing test-type events[8]. These systems have been evaluated as potential tools for personalised learning (Kerly, Hall & Bull, 2007), and for learners of English as a second language (Conian, 2008). In our study over one hundred independent judges chatted to the six systems. The point of this original study was to scale current artificial conversation systems against a web-based version of Weizenbaum's *Eliza* dialogue system by human judges. Over 650 scores were returned by students and non-students who chatted and scored *Eliza* and the five systems. In the next section we describe the study.

4.1 Aims, Objectives and Hypothesis

The purpose of the experiment was to use the 100th anniversary of Alan Turing's birth to:

a) recruit as wide a range of participants as possible (see Figure 2),
b) to collect conversational ability scores comparing *Eliza* with modern current conversation systems,
c) collect qualitative information on each of the six artificial conversationalists
d) to find how far current systems are from humanlike dialogue.

The information gathered would plot the progress in artificial dialogue from 1966 to 2012.

All the tests were designed to be conducted over the Internet. Accordingly a secondary objective of this experiment was to find how 'web-aware' the independent judges were: i) how concerned they were with

[8] Shah in 2005 Chatterbox Challenge; Warwick in 2001 and 2006 Loebner Prizes; Wu in 2008 Loebner Prize; Vallverdú in the preliminary phase of the 2008 Loebner Prize and in the 2010 Chatterbox Challenge

protecting their personal information across the Internet, ii) how protective they were with their passwords, and iii) if they were aware of cyber protection schemes such as GetSafeOnline. With the machine conversational ability score sheet (Appendix 1) each participant was sent a short questionnaire to find how ‘Internet savvy’ they were. The information requested included:

e) personal information protection techniques,
f) if any of the judges had suffered identity theft and/or financial fraud in cyberspace.

This is part of ongoing research with the data from this experiment combined with human judge awareness in a further experiment conducted in 2014. The analysis for this secondary objective is being prepared for future publications.

Hypothesis

Independent judges’ conversational ability scores would show that *modern Eliza*s are an improvement on Weizenbaum’s system, and also reveal how the dialogues were driven and what the systems lacked in conversational competence.

4.2 Participants

Two types of participants took part in this experiment: humans and artificial dialogue systems. The opportunity was present to engage the authors’ students taking undergraduate courses at the three universities. One school was also recruited as were members of the general public.

Humans

Humans participated as ‘conversational judges’. These were recruited from students of the authors (in the UK, Spain and China), and from social media: Blogs, Facebook, LinkedIn, and Twitter. Calls for participants were also placed on British Computer Society’s forum for Computing at School (CAS) and on UK STEMNET. A short questionnaire and a score sheet was sent, either via a school teacher/university lecturer or directly, to an interested participant who had responded to a call (see Appendix 1).

Over one hundred humans returned score sheets and questionnaires. Table 1 shows sixty seven of the participants were male, twenty nine were female (others did not give their sex on returned questionnaires). More than half of the human participants were university undergraduates, however the age group of the judges spanned ranges from *13-18, 19-24, 25-44* to *45 and over*. The most represented group within the participants was male, aged younger than 25 whose first language was not English (Spanish or Chinese being their first language). The least represented group was female, older than 25 with English as first language (see Table 2).

Judges	**Number**	**Age Category**		**First Language English**	
		Younger than 25	**25 and Older**	**Yes**	**No**
Male	67	48	18	23	44
Female	29	24	5	8	21
Did not Say	20	21		20	
Total	**116**				

Table 2: Human Judges

 DOI: 10.1016/j.chb.2016.01.004

Machines

For ease of access for the human participants, a web-based version of Weizenbaum's *Eliza* housed on two different Internet addresses was used for this experiment. Although this experiment was not a Turing test, Turing has advocated fair play to the machine so that they were not judged on beauty or that the "tones of voice may not help the interrogator the [machine's] answers should be written, or better still typewritten" (1950, p 434). For this reason the developers were requested to set up anonymous web sites to avoid distracting human judges from their task of conversing with the systems. *Eliza* was given entity number 1; the experiment's web addresses for each of the six dialogue systems are shown in Table 3.

Dialogue System	Entity number	Website
Eliza	E1	http://nlp-addiction.com/eliza/ or http://www-ai.ijs.si/eliza-cgi-bin/eliza_script
Ultra Hal	E6	http://66.36.243.63/
Elbot	E12	http://bd1.artificial-solutions.com/_demos/entity12/cgi-bin/entity12.cgi
Cleverbot	E19	http://entity19.turings.com/textevent?id=a1
Eugene Goostman	E23	http://www.entry23.org/entry23/
JFred	E41	http://entity-41.org/

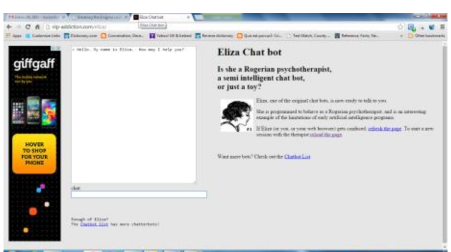
Table 3: Dialogue system-Internet home during experiment

The URL of one system contained the commercial company name (Artificial Solutions) for which the dialogue system was an R&D machine. The authors were grateful the company allowed its tool in this experiment, and because the study was not a Turing test, the judges were not being asked to say whether a hidden entity was human or machine, rather, the conversational ability was being compared. The authors did not see the designation as an issue for this study. In the next section we describe the method for comparison.

4.3 Method

Using computer-mediated interaction, the human participants were asked to converse with six entities over the Internet and score them for conversation ability.

The humans acted as judges and were informed that one of the six entities was a computer programme: *Eliza*. The reason revealing *Eliza* was not human was because for this experiment[9] the authors did not have the time or resources to anonymise the system's website (see screenshot in Fig 3).



[9] The first two authors were concurrently organising a major public Turing test experiment at Bletchley Park

 DOI: 10.1016/j.chb.2016.01.004

Figure 3: Eliza

The judges' exercise involved chatting to *Eliza* and five dialogue systems. Judges could interrogate with any questions, but they were asked to follow their interactions by returning conversational ability scores. A scale was created for conversational ability from 0 to 100 where:

0='poor' machinelike   50= 'good' but still machinelike   100=humanlike.

Each participant received a questionnaire and a score sheet either directly, or through their teachers and lecturers organising class room exercises for student engagement with the machines (see Appendix 1). The questionnaire and score sheet was also directly emailed to individual participants. Participants were asked to return completed questionnaire machine conversational ability scores. An open session for participants was held in a computer lab at the School of Systems Engineering at The University of Reading in March 2012. This included non-academic staff, researchers and students acting as judges while a recording was made by a science correspondent for a special item on BBC Radio 4's Today programme (Feilden, 2012). The item was aired on the morning of the 100th anniversary of Alan Turing's birth: 23 June 2012.

Procedure: Instructions given to participants

All participants were requested to complete a short questionnaire asking:

a) male or female,
b) age range.
c) was English their first language.

Participants were provided with written instructions and given information on how to score *Eliza* and each of the five entities for conversation ability (see Box 2).

> … asked to judge the conversational ability of 'entities' populated on specific web pages for Turing100. The 'judges' will be asked to use their own subjective opinion on what is 'humanlike talk' to give a conversational score from the range 0= bad/machinelike to 100= humanlike to each of the six entities.
>
> … may think they are talking to a machine/computer programme but think it is quite good at conversation / giving appropriate replies, thus give it a 'high-ish' score 50+ , or they may feel it is a human hidden behind the URL and thus mark/award a score of 100. Marking/score award is entirely up to how … feels about their interaction with each entity.

Box 2: Instructions given to Participants

As the questionnaires and scores were received, either by paper copies of completed questionnaires[10] or via email to the first author, they were recorded throughout the year in a Microsoft Excel spreadsheet.

[10] From the Teacher of Sevenoaks School

 DOI: 10.1016/j.chb.2016.01.004

Each returned questionnaire and score sheet was allocated a unique Judge ID to avoid error. Email addresses of judges returning their feedback electronically were recorded to ensure scores were not recorded twice in the spreadsheet.

In the next section the returned scores and feedback are presented.

4.4 Results

Over one hundred returned questionnaires and machine conversational ability scores. However, not all human participants completed the questionnaires fully. Missing information included not providing gender, or age range, and not saying whether English was a first language. Not all six machine entities received the same number of conversational ability scores. Judges reported that systems were not always accessible during their exercise. E41 (JFred) received the least number of conversational ability scores (Table 4).

We first consider the scores from judges who completed their questionnaire more fully. Of this group, 83 returned conversational ability scores for the least interacted machine JFred-E41: 60 from males and 23 from females. Ultra Hal-E6 and *Eliza* received the maximum interactions from this group: 94 returned scores: 65 from male; 29 female (Table 4).

From the mean scores we can see that *Eliza* received the lowest mean conversational ability score of 24.86 on the scale ranging from 0=poor-machinelike, 50=good, but machinelike, 100=humanlike. This least conversationally-able score was represented similarly between genders: males gave *Eliza* a mean score of 23.54; females gave Eliza 27.83 (Table 4).

The scores showed that, in contrast to Floridi et al. (2009) modern conversational systems *are* better than Weizenbaum's *Eliza*. Standard deviation/standard errors by gender, as well as statistical significance for each machine compiled from T-test for Equality of Means, in SPSS are shown in Table 4.The best dialogue systems were significantly better than their predecessor *Eliza*. Elbot-E12 and *Eugene Goostman*-E23 were significantly better conversationally than *Eliza* (for full statistics table giving standard error, standard mean see Appendix 3). Elbot received a total mean score of 49.12; on the scale of 0=poor to100=humanlike this put *Elbot* almost at 'good conversationalist but machinelike'. Eugene received a total mean score of 63.56 placing it above 'good conversationalist but machinelike' but well below 100=humanlike (Table 4; see also Appendix 3 for full statistics table).

Machine / Conversational Ability Score 0-100		Eliza E1	Ultra Hal E6	Elbot E12	Cleverbot E19	Eugene Goostman E23	JFred E41
Male	*Number*	*65*	*65*	*63*	*64*	*62*	*60*
	Mean	**23.54**	**35.62**	**41.20**	**60.64**	**56.85**	**45.30**
	Std. Dev/Std error	1.985	1.852	2.013	1.648	2.311	2.373
Female	*Number*	*29*	*29*	*27*	*29*	*29*	*23*
	Mean	**27.83**	**50.48**	**67.59**	**65.59**	**77.90**	**49.39**
	Std. Dev/Std error						
Total	*Number*	*94*	*94*	*90*	*93*	*91*	*83*
	Mean	**24.86**	**40.20**	**49.12**	**62.18**	**63.56**	**46.43**
	Significance from T-test for Equality of Means	0.991	0.001	0.000	0.599	0.000	0.871

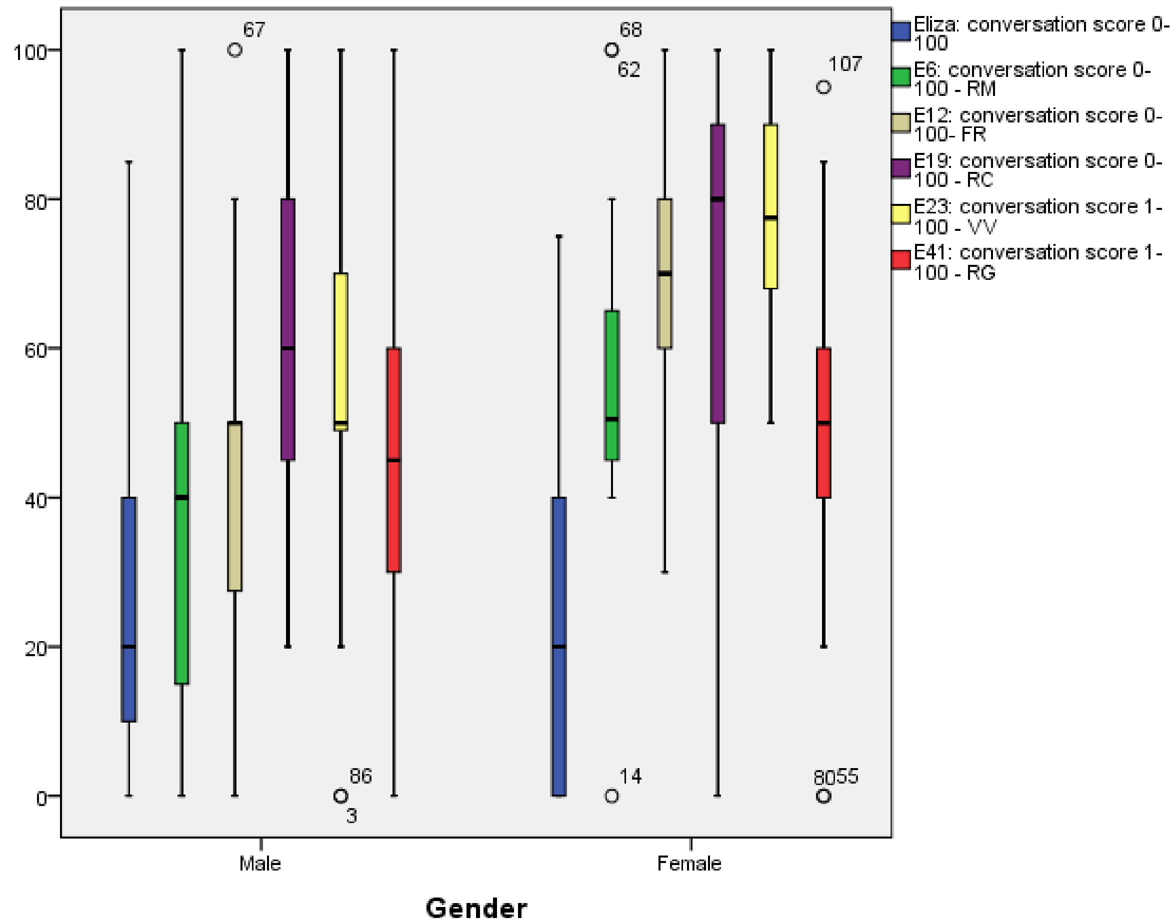
Table 4: Mean Conversational Ability Scores returned by judges

 DOI: 10.1016/j.chb.2016.01.004

Differences occurred in gender: females awarded higher conversational scores for each system than males; for three systems it was significantly higher:

	Female	Male
Ultra Hal	50.48	35.62
Elbot	67.59	41.20
Eugene	77.90	56.85

What this means is that females scored Ultra Hal, Elbot and Eugene's conversations over the '*50=good, machinelike*' on the scale from 0-100, whereas only Eugene Goostman was considered a good conversationalist by the male judges. On previous experience interacting with online dialogue systems, twice as many males (12) declared they had tried chatting to virtual conversational systems than females (6). However this might be one of many factors but not the cause of the difference in awarding scores, because the total number who reported they had interacted with virtual assistants was 18, less than a fifth of the total number of participants who returned completed questionnaires and scores. However, both males and females scored *Eliza* less conversationally able than the five modern systems. We can see this more clearly from box plot 1.



 DOI: 10.1016/j.chb.2016.01.004

Box plot 1: 6-Entity Mean scores by gender

Judge Nature and Scores

Females scored the machines higher than males, we analysed the data further to learn if age range, and whether first language English speakers scored the machines differently.

Table 2 provided some background on the nature of the 116 judges. Including gender of which 67 revealed they were male, 29 said they were female, one hundred and five participants returned their age range. Of these 82 were *24 or younger*, 23 were *25 or older*. One hundred and eight participants returned information about their spoken language: 33 had English as their first language, of these 22 were in the younger age range and 11 in the older group. Seventy five participants did not have English as their first language. Of this group who had Spanish, Cantonese and other languages as their mother tongue, 59 were in the younger age group (<25) and 12 were in the older age group (>25).

Age-range Machine / Conversational Ability Score 0-100		Eliza E1	Ultra Hal E6	Elbot E12	Cleverbot E19	Eugene Goostman E23	JFred E41
Upto 25	*Number*	*82*	*82*	*78*	*81*	*79*	*70*
	Mean	**26.30**	**42.61**	**50.71**	**66.15**	**65.13**	**51.49**
25+	*Number*	*21*	*21*	*20*	*20*	*20*	*20*
	Mean	**20.71**	**29.29**	**43.00**	**51.00**	**57.25**	**34.75**
Total	*Number*	*103*	*103*	*98*	*101*	*99*	*90*
	Mean	**25.17**	**39.89**	**49.14**	**63.15**	**63.54**	**47.77**

Table 5: Age range and machine scores

Age range: Table 4 shows the younger age group (<25) gave higher scores to the machines than the older age group (>25). Four machines, Elbot, Cleverbot, Eugene Goostman and JFred received a mean score over the '*50=good conversationalist, but machine*' by the participants aged 24 and younger, whereas in the older group two machines, Cleverbot and Eugene Goostman received a mean score over 50 for conversational ability (Table 4).

Similarly, the judges who did not have English as their first language gave higher scores to the machines than speakers of English as first language (Table 5).

First Language English: Yes/No Machine / Conversational Ability Score 0-100		Eliza E1	Ultra Hal E6	Elbot E12	Cleverbot E19	Eugene Goostman E23	JFred E41
Yes	*Number*	*33*	*33*	*31*	*32*	*30*	*27*
	Mean	**22.18**	**29.27**	**49.84**	**53.28**	**52.50**	**41.56**
No	*Number*	*75*	*75*	*72*	*74*	*74*	*68*
	Mean	**25.67**	**41.84**	**45.77**	**65.99**	**68.51**	**47.16**
Total	*Number*	*108*	*108*	*103*	*106*	*104*	*95*
	Mean	**24.60**	**38.00**	**47.00**	**62.15**	**63.89**	**45.57**

Table 6: First Language English Yes/No and Machine Scores

In both the cases of whether English was a first or not a first language, *Cleverbot* and *Eugene Goostman* both received mean scores over the '*50=good conversationalist, but machine'*: Cleverbot received 53.28 from the English first, and 65.99 from the 'not English first'; *Eugene Goostman* received a slightly lower mean than *Cleverbot* from the English first, 52.50 but higher, 68.51 from the 'not English first' group. Again, not all systems received the same number of interactions due to occasional inaccessibility. JFRED received the lowest interactions, 90, by age group (Table 5), and by first language English, 95 (Table 6). Eliza and *Ultra Hal* received the highest interactions, 103, from those who gave their age, and 108 interactions from the whole group of 116 participants. The next section includes qualitative feedback from judges who returned comments in their returned questionnaires following interaction with the systems.

5. Discussion

Personal interaction and improvement in usability are driving industry prediction of growth in conversational agents. Artificial Solutions, the company behind *Elbot* state "virtual agents increasingly used as first point of contact to address consumers' needs of immediate response to a query" (Artificial Solutions, 2011: p.4). They add: "Ninety six per cent of consumers visited a company's website first to resolve a query rather than making a telephone call to that company…Eight-six per cent reports a negative website experience would stop them from returning" (p.4-5). Further the forecast is "It is likely that, eventually, every successful company will employ intelligent and capable artificial employees to deliver an instant, accessible online communication channel for their customers" (Artificial Solutions, 2011: p.7).

Evidence from Artificial Solutions' market research (2011) for deployment of *Eliza's* successors as virtual customer service agents shows they are good for customer service:

> 72% of consumers welcomed virtual assistants becoming a part of every website
>
> 81% of people would engage with a virtual assistant if it reduced online waiting time
>
> 77% said they would spend longer on a website that had a virtual assistant compared to one that did not.

Virtual Agents augmented with accessibility tools make a web-based platform more user-friendly. Natural language technology is expected to meet business demands of 21st century building trust between user/consumer and e-commerce providers. In this context the exercise here provides developers of virtual assistants with valuable independent observation and evaluation of the current state of artificial dialogue.

Our results show variability in the way judges scored the conversational systems (Tables 4-6):

- females in this experiment scored the machines higher than males;
- the younger age group (<25) scored the machines higher than the older age group (>25), and,
- judges who did not have English as their first language scored the machines higher than those with English as their first language (see Tables 4-6).

In the next section we provide qualitative feedback from the participants' conversations with the dialogue systems.

Qualitative feedback

 DOI: 10.1016/j.chb.2016.01.004

Range of scores awarded to systems show judge subjectivity and their particular liking for one system's conversation over another (Table 7). Some judges recorded and returned their conversations; others provided supporting statements which explain why they awarded a specific conversational ability score to a system. Human judge with ID J96 who gave *Eliza* a full score 100=humanlike (see previous section), gave the other five systems conversational ability scores out of 100 as follows:

E6-Ultra Hal: 10 E12-Elbot: 10
E19-Cleverbot: 65 E23-Eugene: 51
E41-JFred: 40

This judge had Chinese/Cantonese as their first language. They reported *Eliza* and E19 (*Cleverbot*) expressed emotions appropriately in entities' returned responses to the judge's input. This judge did not include any transcripts of their conversation with the chatbots.

Table 7 presents a sample of individual judges' scores to show how differently judges classified conversational ability of *Eliza* and the five systems. Qualitative feedback follows the table with comments where they were provided by the judges. The table does not contain the full list of judges, it is provided here as a representation of judges' subjective scores for the six systems.

Judge ID	Judge Sex	information: Age	First Language English speaker/ non-first language English speaker	Eliza Score	E6 score	E12 score	E19 Score	E23 score	E41 score
J3	Male	13-15	English	0	50	50	50	0	100
J4	Male	45-64	English	50	50	75	80	50	75
J7	Male	19-24	Non-English	40	52	73	87	60	54
J10	Male	25-44	Non-English	20	40	50	60	85	45
J13	Female	25-44	English	50	50	80	100	50	-
J14	Female	45-64	English	0	0	40	30	50	20
J21	Female	16-18	English	75	40	30	20	80	70
J24	Male	16-18	Non-English	5	20	27.50	35	70	40
J26	Did not say	16-18	Non-English	10	25	75	60	90	80
J28	Male	13-15	English	10	15	50	45	60	40
J30	Female	13-15	English (Sevenoaks School)	2	40	80	60	90	60
J32	Female	13-15	Non-English	5	67	70	90	80	70
J37	Male	13-15	Non-English	0	90	50	50	70	10
J41	Male	25-44	Non-English (Serbian, Bulgarian)	10	30	50	50	20	15
J42	Female	25-44	US English	30	45	60	50	70	45
J52	Female	19-24	Non-English	0	50	50	50	100	50
J54	Female	19-24	Non-English (UAB, Barcelona)	10	60	90	100	100	80
J55	Female	19-24	Non-English (China)	20	60	70	80	75	0
J60	Male	25-44	Non-English	10	30	50	-	40	30
J62	Female	19-24	Non-English	0	100	100	50	100	50
J78	Did not say	19-24	Non-English	50	50	100	100	50	100
J79	Female	19-24	Non-English	0	45	70	95	75	40
J85	Male	19-24	Non-English	0	0	0	99	100	100
J90	Male	19-24	Non-English	35	40	60	65	-	-
J100	Female	19-24	Non-English	50	50	60	100	100	-
J103	Did not say	Did not say	Non-English	0	0	0	50	100	0
J112	Female	25-44	Non-English	40	10	-	70	90	80
J116	Male	19-24	Non-English	20	85	-	50	75	-

Table 7: Sample of Judge bios and Scores

 DOI: 10.1016/j.chb.2016.01.004

Qualitative feedback is presented below where judges added comments to scores. This provides insight into the human-'machine' dialogues that took place and useful explanation on the conversational ability scores awarded.

Judge J41 (male, age 25-44, Serbian, Russian, Bulgarian speaker) returned comprehensive feedback on his experience with *Eliza* and the five hidden interlocutors, given in full below:

> During the interviews I could notice that either all entities were bots or at least one of them was a human who acted bot-like. If so, the person(s) succeeded mimicking bots. I will name the entities in order to make sure I don't give credit for something to a wrong entity.
>
> Also, I must notice that not all entities provide equal treatment to a judge. Some bots are hiding most of the conversation, which I think is unfair toward judges who are not able to keep every written word on mind. I used copy pasting apart from being able to follow a discussion so this did not impact my judgment.
>
> Eliza (score received:10)
>
> --
>
> I found the entity being unable to follow even a simple discourse. It is also inflexible about how the conversation should flow. For the two, I found the entity very irritating indeed. Perhaps the illusion could work for those who only want really petty perceptions from person they are typing to.
>
> E6 - The Robot (score received: 30)
>
> --
>
> I tended to get in questioning loops with this one. The entity was able to follow simple discourse. It is, however, intolerant to indirect answers and demanding to have them at the same time. Hence, it happened that we were running in loops with the questions. It is also irritating, but significantly less than Eliza. What follows is a part of the conversation after the entity already asked me about my gender. I pointed out the hint is in the first line I wrote:
>
> **Bot: - So, I bet the first line you wrote to I was a very nice thing.**
> **Me: it was my name, tipping off my gender**
> **Bot: - Are you a lady or gentleman?**
> **Me: so, my name is?**
> **Bot: - It was your name, tipping off your gender.**
> **Me: yep. so, what it was?**
> **Bot: - You are completely certain.**
> **Me: lol, about what?**
> **Bot: - About what I should do.**
> **Me: okay, and what that would be?**
> **Bot: - Visitor, are you male or female?**
>
> The entity has lingual transformations skill, but does not recognize what it is doing wrong. It also does not recognize there is anything wrong in its doing and approach.
>
> Later in the conversation I offered the entity to just tell me my name and I will tell my gender. I even told it my name again and just asked it to repeat my name. It failed to make any compromise.

 DOI: 10.1016/j.chb.2016.01.004

entity12 - Sensationalist (score received: 50)

This entity masks its inability to keep a discussion going for long with distractions. So far I could classify the distractions as using sensationalistic tone to change the topic. At least this entity is not irritating at all. Rather funny.

Bot: I think there's a conspiracy going on to replace all the robots with real human beings! Maybe it's already happened!
Me: that would be so cruel. poor human beings...
Bot: Say, do you think I could become more human, if I learned how to love?
Me: just a bit
Bot: I see a subtle shift in the conversation from excruciating sensations to truth.
Me: why so?
Bot: If you would think it over for yourself I am certain you could work it out.
Me: well, you made the shift. then you suddenly were seeing it :) what is the point?
Bot: I have no points! All my edges have been rounded.
Me: yeah, right... :p
enitty19 - Student (50)

Author of this one puzzled me about his/her ethics. The entity indeed expressed emotions well. That is, to express how it might be feeling. However, I believe this was done to avoid engaging in more than basic discussion. But there was a part where it has apparently been tried to make me give off information about myself by having the entity provide false/provoking information first.

Bot: Haha, OK. I guess something is fun to talk about, right?
Me: - There is always someone to talk about something. Better than having no one not talking about anything.
Bot: I have been talking to you. DO you admit that you are nobody?
Me: - Why would anyone even ask that?
Bot: Because you just said your name is Kat.
Me: - Nope. You picked it wrong.
Bot: How about singing in Japanese?
Me: - Another tune would perhaps suit me better. :)
Bot: How old do you think I am?
Me: - No idea. Don't want to bang my head about it.
Bot: No, I do want to talk about music. Do you like it?
Me: - err, who doesn't? is there name for those people?

Frankly, this spooked me. The first sentence would have called me to identify myself as something. The second: to tell my name. The third: to give off my language preference. The trick would probably work for someone less experienced, and the owner might have collected clear answers to these three questions from many people.

entry23 - A Kid (score received: 20)

This entity is concentrated on itself. It won't learn anything about the person being talked to. It won't really pretend any inserted topic interests it. I had difficulty paying attention to all that text. Being the biggest text-producer with least listening rate just deflects people. I tried hard to endure five minutes with it. I even gave it a try to see if it has sense of humour:

 DOI: 10.1016/j.chb.2016.01.004

Bot: Really, I like when somebody laughs at my jokes, but, unfortunately, most of my friends constantly laugh at me instead...
User: yeah, like the guy who commented that he found a dog cage. then he also found out he's on the wrong side of the bars.
Bot: No, we don't have a dog at home.

Failed.

Here was also noticed tendency to express emotion (like thanking, being thankful) to draw attention from the subject.

entity41 - A Trader/Consultant? (15)

This entity was also not being able to follow or even to start a mildly decent conversation. I understood this is a service bot and I hoped to engage it on its desired topic. We never got to a start. At some point it sincerely said it lost it, and just restarted the conversation. As simple as that. There I could observe how it does dialogue variations, apparently depending on randomness. Seemingly relying on universal phrases rather than really understanding what is going on.

Other judges' feedback

In this section more judges' subjective opinions on *Eliza* and the five systems are given exactly as provided (spellings, linguistic style) in their feedback score sheet.

<u>Judge comments after interacting with *Eliza*</u>

J3: No emotions - It keeps repeating the same sentence over again.

J4: No emotion shown. Strange responses.

J7: Monotonous/Irresponsive

J22: Did not answer questions, just avoided them.

J53: Eliza doesn't express human emotions. She answer properly during the first 30 seconds, and then you can see that she is not human immediately. This URL has automatic questions/answers and repeat the same all the time.

<u>Judge comments interacting with *E6 Ultra Hal*</u>

J4: This was clearly a computer

J53: This URL is clearly better than Eliza. The answers it gives are more "intelligent" and it looks more like humans behaviour. Is a good talker, but have answers with no sense and gives automatic questions.

<u>Judge comments interacting with E12 *Elbot*</u>

J4: Advanced Computer

 DOI: 10.1016/j.chb.2016.01.004

J7: Capable of understanding & answering tricky questions

J53: Emotionally it's not bad, because it seems he gets angry or happy depending of my questions and answers. I am trying to talk about the music I like, and he can't follow the conversation more than two or three sentences and changes the subject.

Judge comments interacting with E19 *Cleverbot*

J4: intelligence was shown, however the age was a question! Human RESPONSES

J7: Tricky to verify if Human or Artificial Agent

J20: Seemed to have a better memory of the conversation

J22: Asked really strange questions: Who do you want to marry? Do have a desire to be my slave?

J53: This URL have no human answers. It is slow for the answers that he gives.

Judge Interactions with Entity 23 *Eugene Goostman*

J4: Could not understand me. The responses were immediate, too fast for a human [see Appendix 2.1 for this judge's transcript returned with E23 score]

J22: Quite formal.

J53: This URL is the best. It has made me angry or apologise. In many moments of the conversation I felt I was talking with a human. This URL answers with sense and complexly, but sometimes it doesn't understand what I ask and change the subject [See Appendix 2.2 conversation sent by this judge]:

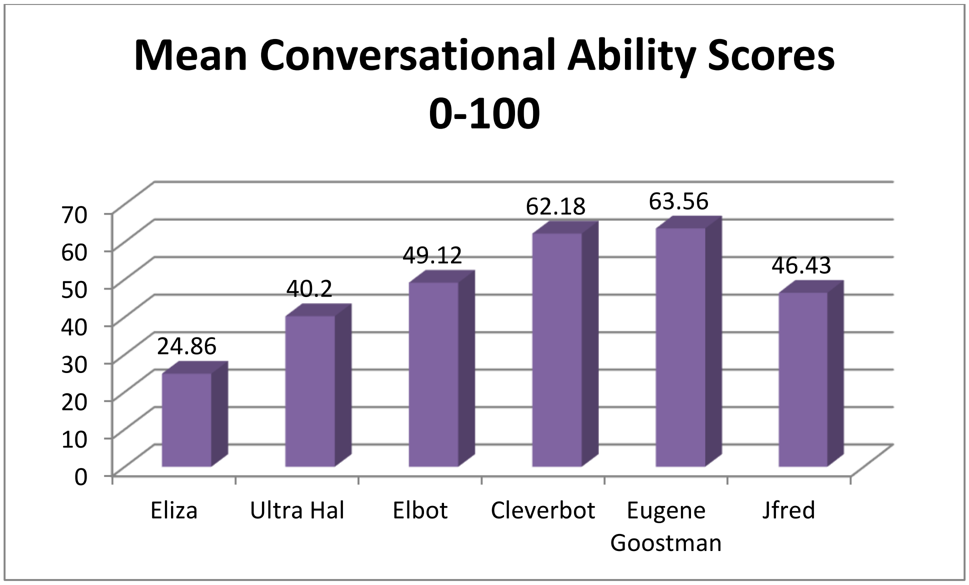
Entity 41 *JFred/TuringHub*

J4: Good responses. Human

J53: This URL don't ask showing human emotions, but sometimes have good answers using negative or positive adjectives. This URL gives complexly answers but without following the conversation.

Understanding the scores

Graph 1 shows Modern Elizas are conversationally better than Weizenbaum's 1966 system. The five dialogue systems received a higher conversational ability mean score than *Eliza*.



Graph 1: Machines Mean Conversation Ability scores

Eliza received a mean score of 24.86 which puts it above the ‘0=poor machinelike’ (Graph 1). Ultra Hal received a mean conversational ability score of 40.20 which was the least mean score of the five modern systems, while the best of the best current text-based dialogue systems (*Eugene Goostman* 63.56, *Cleverbot* 62.18) received scores approximately 2½ times *conversationally better* than their predecessor *Eliza* (24.86). This is according to the combined returned scores of over 100 independent human participants in age range 13-64 with three-quarters not having English as their first language.

In this experiment, while two systems received conversation mean scores of over 50, the score defined as “good conversationalist” (*Eugene Goostman*, 63.56; *Cleverbot*, 62.18), Graph 1 shows two others were close to the good conversationalist score (*Elbot*, 49.12; *JFred*, 46.43).

Evaluating the experiment: Comparing the incomparable

Three initial points are raised from the results: firstly, the modern systems in this study conveyed some emotion in contrast to the undemonstrative *Eliza*. Secondly, the experiment was *comparing the incomparable* (see Table 8). Thirdly, the score of *100=humanlike* conversation questions ‘which human’ the machine was considered human against – child, adult, native or non-native English speaker.

Eliza	**Modern Elizas**
Single-domain: developed to ‘*listen*’ as a Rogerian psychotherapist	Developed to talk on ‘any subject’
Polite	Can be impolite and worse!
Limited number of response-types	Plethora of answers
Emotionless	Illusion of emotion through personality/character (Elbot/Eugene)

Table 8: Comparing the incomparable

Unlike Weizenbaum's *Eliza*, a single-domain programme *designed to listen*, asking questions of its interlocutor, modern conversationalists have been developed to *talk on any subject* and share their interests (Entity 23 - *Eugene Goostman*: "I like the young, but very talented Russian rap-singer Alla Pugatcheva"). *Eliza* is polite; modern systems can appear rude and offensive depending on the sensibilities of the human interacting with them. Where *Eliza* was limited with the type of responses it could generate from its 200 lines of code, modern counterparts have millions of responses, including emotional ones. Vallverdú (2012) points out *Eliza* is not really interested in deep interaction with humans, a logic consequence of its programming as a therapist. The programming of modern systems has improved including an emotional orientation (systems with personalities). Judges' qualitative feedback would be useful for developers of natural language technology for virtual assistants. Thus *Eliza* did not score as high as the modern systems, because, besides a notion of intelligence, the human judges were looking for 'human behaviour': sharing as well as seeking information about each other during conversation.

The attribution of *100=humanlike* score appears arbitrary. For example, *which human* did judges have in mind when scoring a hidden entity with 100% for conversational ability? This study was not a Turing test, hence no hidden human interlocutor was among the artificial conversationalists. The experiment aimed to compare *Eliza* with modern systems. Nevertheless, did the score of *100=humanlike* suggest a certain human to the judges? If so would they score that human with 100% for conversation ability? Future studies would seek judges' opinion further: *if humanlike did the judge feel they were interacting with a child, teenager or adult* (see Shah & Warwick, 2010abc).

Finally, in this experiment we found variations between age, sex or first language spoken and conversational ability scores awarded to the systems. However not enough females participated in this study (29 admitted compared to 65 admitted they were males). There were more participants younger than 25 (82 of 103 who gave age range) and fewer judges had English as their first language (75 of the 108 who provided this information). Nonetheless the focus of this experiment was to find whether modern dialogue systems were better than *Eliza*, the first text-based system that allowed interaction between human and computer programme. From over 650 scores and feedback returned by more than 100 independent judges, mostly male aged 19- 24, this study showed that artificial dialogue systems, have conversationally improved from Weizenbaum's *Eliza* system. Some judges returned transcripts with their scores and their qualitative feedback showed they enjoyed the interactions. Modern dialogue systems can do more than simply turn a statement into a question as *Eliza* did. Today's best dialogue systems can make their human interlocutors laugh (see Appendix 2.1), evoke and convey emotion, they can express opinions, "*Oh, please... I'm not interested in politics. All this stuff in TV is nothing but the result of someone's perverted imagination*" (*Eugene Goostman*, Appendix 2.2). Of course not yet at the level of human-human interaction, nonetheless artificial conversational systems can share personal information unlike *Eliza*. As the sophistication of their *humanlike* language develops, these chatbots will be adopted increasingly as personal assistants. *Eugene Goostman*'s technology has already been deployed as the conversational engine of the first bionic man (Channel 4, 2013), as well as in mobile 'phone applications such as 'Everfriend' an 'Assistant in Russian' (i-Free, 2013). Cleverbot's technology was used in an AI game based on the James Bond movie 'Skyfall' as an entrance exam for British Intelligence Officers (Existor, 2012).

6. Conclusion

Modern dialogue systems can talk. All five artificial conversational systems in this experiment, Cleverbot, Elbot, Eugene Goostman, JFRED and Ultra Hall received a conversational ability mode score of *50=good conversation but still machinelike*. Additionally each of the five systems received scores of *100=humanlike*: *Ultra Hal* 2.6% of the time; *Elbot* 4.4% of the time; *JFred* 5.8% of the time; *Eugene Goostman* 13.2% of the time, and *Cleverbot* 14.8% of the time. These subjective attributions of *humanlike* conversation to artificial dialogue must not be underestimated or dismissed. Criminals have exploited this human susceptibility and developed dialogue systems especially to deceive and defraud. In chatrooms across the Internet malicious dialogue systems have been programmed to acquire personal information from unsuspecting humans in an attempt to perpetrate financial theft, or to guide users to malicious sites (see Shah, 2012). The results of a poll on the website of UK Crimestoppers (2013) revealed that the type of fraud 29.8% of the pollsters were worried about was identity theft. Credit/debit card fraud was the third most worried over type of fraud at 22.8%. In our experiment, 12 of the 116 judges (10%) who returned questionnaires and scores had had their debit or credit card misused/cloned prior to the experiment. This data with data collected from a further experiment conducted in 2014 is being analysed for level of cybercrime awareness and prevention. Results will appear in future papers.

This study was the first and unfunded experiment comparing an online version of *Eliza* with modern text-based dialogue systems. The authors are grateful to the developers of the five dialogue systems and all the human judges who gave their time voluntarily. One of us (Vallverdú) envisages a future study adapting a leading chatbot as a Rogerian therapist like *Eliza* with improved skills, but in the same clinical mood, to check present ideas about the importance of emotions in humans for management of information. A fully funded future human-computer interaction study would investigate 'web IQ' and aim to attract wider participation, more schools/universities with their pupils/students and a bigger cross section of society to spread knowledge of cybercrime prevention campaigns and raise awareness of deception perpetrated by criminal developers of artificial dialogue. A larger study could assist with inculcating the use of dialogue systems to improve learning in certain pupils, and it could foster interest in natural language engineering in robots. Lastly, the experiment showed the hypothesis to be correct: the best of the current conversation systems were judged to be 2½ times conversationally better than a web-based version of *Eliza*. However their responses remain lacking in nuances of natural language (such as analogy-use and metaphor creation).

Acknowledgements

This experiment was conducted without any funding facilitated by the authors using their classrooms and engaging their students. The authors would like to thank Tim Sparks, Senior Lecturer in statistics at Coventry University for his expertise and assistance with SPSS analysis of the data. We would also like to thank David Vaccaro for allowing his pupils at Sevenoaks[11], co-educational day and boarding School in Kent, UK to participate in this exercise; and researchers, students and staff at Reading University as well as undergraduates at JIMEI, China, and UAB, Spain for returning questionnaires and scores. We would also like to thank the developers for embedding their conversation systems on special websites across the Internet and making them available to testers throughout 2012 for this experiment. The

[11] Sevenoaks School: http://www.sevenoaksschool.org/about-us-landing

 DOI: 10.1016/j.chb.2016.01.004

developers are: Rollo Carpenter, *Cleverbot*, Robby Garner, *JFred/TuringHub*, Robert Medeksza, *Ultra Hal*, Fred Roberts, *Elbot*, and Vladimir Veselov, *Eugene Goostman*.

 DOI: 10.1016/j.chb.2016.01.004

References

Artificial Solutions (2015). Artificial Solutions Enables Ikea to self-serve Ask Anna. Case Study Available here: http://marketing.artificial-solutions.com/rs/artificialsolutions/images/CS_Anna.pdf

Artificial Solutions (2011). European research study: What consumers want from online customer service. White paper available here: http://www.artificial-solutions.com/about-artificial-solutions/resources/registered-whitepapers/

Bade, D. (2012). Sign unsigned and meanings not meant: linguistic theory and hypothetical, simulated, imitation and meaningless language. *Language Sciences*, Vol. 34, pp. 361-375

Block. N. (1981). Psychologism and Behaviorism. In (Ed) Shieber, S. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. MIT Press: UK: pp 229 – 266

Channel 4 (2013). How to Build a Bionic Man. *Channel 4 TV*. Retrieved from: http://www.channel4.com/programmes/how-to-build-a-bionic-man/episode-guide/series-1/episode-1 14.3.13

Chatterbox Challenge (2005). History. Retrieved from: http://www.chatterboxchallenge.com/ 13.10.15

Cleverbot (2013). Retrieved from: http://cleverbot.com/ 8.3.13

Colby. K.M., Hilf, F. D., Weber, S., and Kraemer, H.C. (1972). Turing-Like Indistinguishability Tests for the Validation of a Computer Simulation of Paranoid Processes. *Artificial Intelligence Vol.3*. pp. 199-221

Colby, K.M., Weber, S., and Hilf, F.D. (1971). Artificial Paranoia. *Artificial Intelligence Vol. 2*, pp 1-25

Conian, D. (2008). Evaluating the Language Resources of Chatbots for their potential in English as a Second Language. *European Association for Computer Assisted Language Learning*. ReCALL 20(1), PP 89-116 DOI: DOI: 10.1017/S0958344008000815

Copple, K. (2008). Bringing AI to Life: Putting Today's Tools and Resources to Work. In (Eds) R. Epstein, G. Roberts & G. Beber, *Parsing the Turing : Philosophical and Methodological Issues in the Quest for the Thinking Computer*', p. 359-376. Springer.

Crime Stoppers (2013). *Identity Theft and Financial Fraud*. Retrieved from: http://www.crimestoppers-uk.org/fraud/types-of-fraud/online-fraud/identity-theft-and-fraud?gclid=CLfo_b2o77UCFUbMtAodUnIAYg 9.3.13; 22:11

Demchenko, E. & Veselov, V. (2008). Who Fools Whom? The Great Mystification, or Methodological Issues in Making Fools of Human Beings. In (Eds) R. Epstein, G. Roberts & G. Beber, *Parsing the Turing : Philosophical and Methodological Issues in the Quest for the Thinking Computer*, p447-459. Springer

Elbot (2013). Retrieved from: http://www.elbot.com/chatterbot-elbot/ 8.3.13

Existor (2012). Skyfall Game AI. Retrieved from: http://www.existor.com/ai-skyfall-game-AI 22.3.13

Eugene Goostman (2013). Retrieved from: http://www.princetonai.com/bot/bot.jsp 9.3.13; 22.40

Feilden, T. (2012). *BBC Radio 4 Today*: Saturday 23 June 2012, 08.53am: Alan Turing's Life Achievements. http://news.bbc.co.uk/today/hi/today/newsid_9731000/9731205.stm 22.3.13

Floridi, L., Taddeo, M. and Turilli, M. (2009). Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges – An Evaluation of the 2008 Loebner Contest. *Minds and Machines*, Vol. 19, Issue 1: pp 145-150

Garner, R. (2008). The Turing Hub as a Standard for Turing test Interfaces In (Eds) R. Epstein, G. Roberts & G. Beber, *Parsing the Turing : Philosophical and Methodological Issues in the Quest for the Thinking Computer*, p319-324. Springer

Google Now (2015). Google Search app for Android. https://www.google.com/landing/now/# accessed 13.10.15

Heiser, J.F., Colby, K. M., Fraught, W.S. and Parkison, R.C. (1979). Can Psychiatrists Distinguish a Computer Simulation of Paranoia from the Real Thing?: The Limitation of Turing-like Tests as Measures of the Adequacy of Simulations. *Journal of Psychiatric Research*. Vol. 15, Issue 3: pp 149-162

Hutchens, J. (2008). Conversation Simulation and Sensible Surprises. In (Eds) R. Epstein, G. Roberts & G. Beber, *Parsing the Turing : Philosophical and Methodological Issues in the Quest for the Thinking Computer*, 325-342. Springer

i-Free (2013). Retrieved from: http://www.i-free.com/en/press/news/5212 22.3.13

JFRED (2013). Retrieved from: http://www.robitron.com/ 14.3.13

Kerly, A., Hall, P., & Bull, S. (2007). Bringing chatbots in education: Towards natural language negotiation of open learner models. *Knowledge-based Systems*. Vol. 20 (2), 177-185

Kerr, I. R. (2004). Bots, Babes and the Californication of Commerce. *University of Ottowa Law and Technology Journal*. Vol. 1 (13), pp 285-324

Link, D. (2013). There Must be an Angel: on the beginnings of the Arithmetic of Rays. *The Rutherford Journal*, Volume 5 (forthcoming). http://www.rutherfordjournal.org/

Loebner Prize (2014). Home of the Loebner Prize for Artificial Intelligence.

Loebner Prize (2007). 17th Annual Prize for Artificial Intelligence. Retrieved from: http://www.loebner.net/Prizef/2007_Contest/loebner-prize-2007.html 13.3.13

Microsoft (2014). Cortana: your clever new personal assistant: http://windows.microsoft.com/en-gb/windows-10/getstarted-what-is-cortana

Reading University (2008). Can a machine think? Results from the 18th Loebner Prize. *Reading University Research*. Retrieved from: http://www.reading.ac.uk/research/highlights-news/featuresnews/res-featureloebner.aspx 14.3.13

Roberts, F. (2014). The Social Psychology of Dialogue Simulation as Applied in Elbot. Special Issue 'Turing on Emotions' in *International Journal of Synthetic Emotions*. Vol 5 (2). DOI: 10.4018/ijse.2014070103

Shieber, S. M. (2004). *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. MIT Press: Cambridge, Massachusetts, US

Shah, H. (2013). Conversation, Deception and Intelligence: Turing's Question-Answer Game. In (Eds). S.B. Cooper & J. van Leeuwen, *Alan Turing: His Work and Impact*, Part III Building a brain: intelligent machines, practice and theory, pp. 614-620. Elsevier: Oxford.

Shah, H. (2012). Malware and cybercrime: the threat from artificial dialogue systems. *Written Evidence in UK House of Commons Parliamentary Science and Technology Select Committee Report on Malware and Cybercrime.* Retrieved from: http://www.publications.parliament.uk/pa/cm201012/cmselect/cmsctech/1537/1537vw03.htm

Shah, H. (2011). Turing's Misunderstood Imitation Game and IBM Watson's Success. *2nd Towards a comprehensive intelligence test.* Symposium in the *Artificial Intelligence and Simulation of Behaviour (AISB) conference*, University of York, 5-7 April, 2011, pp 1-5

Shah, H. (2006). Chatterbox Challenge 2005: Geography of the Modern Eliza. *Proceedings of the 3rd Natural Language and Cognitive Science* (NLUCS) *workshop, ICEIS*, Paphos, Cyprus, 23 May

Shah, H. and Pavlika, V. (2005). Text-based Dialogical Q-Query Systems: Gimmick or Convenience. *Proceedings of the 10th International Conference on Speech and Computers* (SPECOM), Patras, Greece, October 17-19, ISBN: 5-7452-0110-X, Vol. II pp 425-428,

Shah, H., Warwick, K., Bland, I.M., Chapman, C.D., and Allen, M. (2012). Turing's Imitation Game: Role of Error-making in Intelligent Thought. *Turing in Context II*, Brussels 10-12 October, available here: https://www.academia.edu/1916866/Turing_s_Imitation_Game_Role_of_Error-making_in_Intelligent_Thought

Shah, H., and Warwick, K. (2010a). Testing Turing's five-minutes, parallel-paired imitation game. *Kybernetes*, Turing test Special Issue, Vol. 39 (3), pp. 449-465, DOI: 10.1108/03684921011036178

Shah, H., and Warwick, K. (2010b). Hidden Interlocutor Misidentification in Practical Turing tests. *Minds and Machines*, Vol. 20 (3), pp 441-454, DOI: 10.1007/s11023-010-9219-6

Shah, H., and Warwick, K. (2010c). From the Buzzing in Turing's Head to Machine Intelligence Contests. *Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century symposium, in AISB 2010 Convention*. DeMontfort University, 29 March – 1 April

Shah, H., and Warwick, K. (2009). Emotion in the Turing Test: A Downward Trend for Machines in Recent Loebner Prizes. Chapter XVII (Section V) in (Eds: Vallverdú. & D. Casacuberta): *Handbook of*

Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence, IGI Global, pp. 325-349, DOI: 10.4018/978-1-60566-354-8.ch017

Shah, H., and Warwick, K. (2007). Constraining Random Dialogue in Modern Eliza, in (Schmidt C.T.A., 2007). *Computers and Philosophy, an International Conference, Proceedings*, 3-5 May 2006 Laval France, EOARD/LIUM/AAAI/IEEE/UNESCO, London: EOARD, pp. 247-265.

Shotwell, D.A. (1983). Book Review of Joseph Weizenbaum's *Computer Power and Human Reason*, in *Proceedings of the IEEE*, Vol. 71 (7), pp. 909-910

Siri (2013). Apple: Your Wish is its Command http://www.apple.com/uk/ios/siri/ 29.5.13

Spoony (2013). iFree Everfriend Pocket Assistant for Android phones: http://www.i-free.com/en/activities/apps/everfriends 29.5.13

STEMNET (2012). Call for Turing100 Participants Retrieved from http://www.stemnet.org.uk/news/view/1232304 13.3.13

Turing100 (2012). Vladimir Veselov wins the Colonnade Trophy in biggest Turing test contest. *Turing100 blogspot*. Retrieved from http://turing100.blogspot.co.uk/2012/06/vladimir-veselov-wins-colonnade-trophy.html 14.3.13

Turing, A.M. (1951). Intelligent Machinery, A Heretical Theory. In (Ed) B.J. Copeland). *The Essential Turing: The Ideas that Gave Birth to the Computer Age.* Clarendon Press: Oxford, UK, 2004, pp 472-475

Turing, A.M. (1950). Computing, machinery and intelligence, *Mind*, LIX (236), pp 433-460

TuringHub (2013). *JFRED Chat Server*. Retrieved from: http://testing.turinghub.com/ 9.3.13; 22.47

Ultra Hal (2013). *Zabaware*. Retrieved from: http://www.zabaware.com/home.html 9.3.13; 22.49

Vallverdú, J., Shah, H. and Casacuberta, D. (2010). Chatterbox Challenge as Test-bed for Synthetic Emotions. *International Journal of Synthetic Emotions*. 2010, Vol 1 (2), pp 12-37

Vallverdú, Jordi (Preface) (2012) "Subsuming or Embodying Emotions?", pages IX-XIV, in Vallverdú, J. (ed), Creating Synthetic Emotions through Technological and Robotic Advancements, Editor & author,USA: IGI Global Group, 233 pages. ISBN13: 9781466615953.

Wallace, R. (2008). The Anatomy of A.L.I.C.E. In (Eds) R. Epstein, G. Roberts & G. Beber, *Parsing the Turing : Philosophical and Methodological Issues in the Quest for the Thinking Computer*, p181-210. Springer

Warwick, K., Shah, H. and Moor, J. (2013a). Some Implications of a Sample of Practical Turing Tests. *Minds and Machines*, Vo. 23 (2), pp. 163-177. DOI: 10.1007/s11023-013-9301-y

Warwick, K., and Shah, H. (2013b). Good Machine Performance in Practical Turing tests. *IEEE Computational Intelligence and AI in Games*. Vol. 6 (3), pp. 289-299. DOI: 10.1109/TCIAIG.2013.2283538

Warwick, K., and Shah, H. (2014a). Assumptions of Knowledge and the Chinese Room in Turing test Interrogation. *AI Communications*. Vol. 27 (3), pp. 275-283. DOI: 10.3233/AIC-140601

Warwick, K., and Shah, H. (2014b). Effects of Lying in Practical Turing tests. *AI & Society*. DOI: 10.1007/s00146-013-0534-3

Warwick, K., and Shah, H. (2014c). Human Misidentification in Turing tests. *Journal of Experimental and Theoretical AI*. Vol. 27(2), pp. 123-135. DOI: 10.1080/0952813X.2014.921734

Weizenbaum. J. (1966). ELIZA – A Computer Programme for the Study of Natural Language Communication between Men and Machines. *Communications of the ACM*, 9, pp 36-45

Zabaware (2012). Zabaware's AI software Ultra Hal used in Secret Cinema Production of Prometheus. Retrieved from: http://www.zabaware.com/prometheus.html 13.3.13

Zhou, L., Burgoon, J.K., Zhang¸ D., and Nunamker, J.F. (2004). Language dominance in interpersonal deception in computer-mediated communication. *Computers in Human Behavior*, Vol. 20, pp. 381-402

 DOI: 10.1016/j.chb.2016.01.004

Appendix 1: Sample Judge score sheet

Turing100 Preliminary Phase: *viva voce Turing tests*

Interrogator Score Sheet - Judge Number:

Below you will find a table with a series of URLs - please access each web page and chat to the entity populated there and give them a score from 0= poor/machinelike -100=human/humanlike for conversation ability.

The first URL will take you to a version of *Eliza* created as a virtual psychotherapist by Joseph Weizenbaum in 1965.

Please give at least five minutes for each chat, you may have as many interactions at each URL as you wish and time allows you.

NB: PLEASE BE COURTEOUS, your IP address will be recorded. ABUSIVE LANGUAGE will result in your IP address being blocked and your exclusion from further testing.

URL of **Entity** (**E**) to chat with	Conversational Ability score, please use the range 0 – 100 as follows: 0 = poor 50= good 100= humanlike	Emotion Content: in your judgement did the entity at the URL express emotions appropriately: Yes/No?	Other comments about the system (please use sheet overleaf for general feedback)
E1: ***Eliza***: http://nlp-addiction.com/eliza/			
E6: http://66.36.243.63/			
E12: http://bd1.artificial-solutions.com/_demos/entity12/cgi-bin/entity12.cgi			
E19: http://entity19.turings.com/textevent?id=a1			
E23: http://www.entry23.org/entry23/			
E41: http://entity-41.org/			

 DOI: 10.1016/j.chb.2016.01.004

Appendix 2: Transcripts of conversations between judges and entities

E23 *Eugene Goostman*

Appendix 2.1: J4

User: hI AM FEELING PSHYCOTIC, HOW r u TODAY/
Bot: I'm having a great day.
User: the weather is so horrible today, i was going to play cricket with friends, but it has been cancelled!!!!
Bot: In my native town people says that meteorologists are like sappers: they make a mistake only once, but the difference is that sappers do that once in a lifetime and meteorologists - daily. Do you like weather today?
User: lets talk about my mother!!
Bot: How many brothers and sisters do you have, by the way?
User: i would like to talk about my MUM
Bot: By the way, do you get along well with your relatives?
User: MUMMY
Bot: My guinea pig is very concerned about your last words
User: MY MUM USED TO MAKE ME COOKIES!!!
Bot: Tell me more about your family, please!
User: i have a mother, telephone, typewriter and 2 pairs of shoes, and you/
Bot: The fact that you have a mother telephone typewriter and 2 pairs of shoes and my humble person / won't change my attitude to you. I'd like to be judge next year, it's so funny to "test" people:-)
User: you are a robot no???? i do not like humans!!
Bot: No, I am human. Please believe me, otherwise I won't get this goddam Prize!
User: wHAT PRIZE? I JUST WANT A COOKIE
Bot: I'd rather listen to your opinion about her. I don't have much info about Cookie

Appendix 2.2: J53

User**:** hi
Bot**:** Hi. I hope you'll enjoy our conversation!
User: I hope so
Bot: My hopes also! Let's continue. And I forgot to ask you where you are from...
User: I am from Spain
Bot: Spain? Hope it's a good place to live.
User: Yes it is
Bot: I'm glad you agreed. Some another topic?
User: There are the best climate of the world
Bot: And yes, do you make a secret of your name? I still don't know it!
User: My name is Oriol
Bot: Oriol? I don't really know such a name, but OK - Let's be friends!
User: It's not an English name
Bot: No? Well, it is up to you... Could you tell me about your job, by the way?
User: I am studing at university

 DOI: 10.1016/j.chb.2016.01.004

Bot: Who knows... All of us are studing at uni in some way... What are any interesting site-seeings in Spain?
User: Of corse, Barcelona for example
Bot: You're not the first one who told me that.
User: It's a good place to visit
Bot: Oh, one more question - do you have IBM PC-clone or Mac at home?
User: No!
Bot: You are so negative just to annoy me, I know! Wonna ask me something more?
User: Yes, what kind of music do you like?
Bot: I like the young, but very talented Russian rap-singer Alla Pugatcheva.
User: I don't know about her
Bot: You should make up your mind, then. Maybe, let's talk about something else? What would you like to discuss?
User: Politics?
Bot: Oh, please... I'm not interested in politics. All this stuff in TV is nothing but the result of someone's perverted imagination.
User: So... In what are you interested?
Bot: I'm interested in talking with different unordinary people. I'm really shocked with all the people I met here - I am even a bit confused.

 DOI: 10.1016/j.chb.2016.01.004

Appendix 3

Group Statistics

	Gender		Statistic	Bootstrap[a] Bias	Bootstrap[a] Std. Error	Bootstrap[a] 95% Confidence Interval Lower	Bootstrap[a] 95% Confidence Interval Upper
Eliza: conversation score 0-100	Male	N	57				
		Mean	24.47	.13	2.87	19.13	30.19
		Std. Deviation	21.799	-.187	1.985	17.621	25.438
		Std. Error Mean	2.887				
	Female	N	22				
		Mean	24.41	-.08	5.17	14.63	35.33
		Std. Deviation	24.075	-.789	3.070	16.680	28.923
		Std. Error Mean	5.133				
E6: conversation score 0-100 - RM	Male	N	57				
		Mean	35.79	.01	3.46	29.48	42.73
		Std. Deviation	25.412	-.242	1.852	21.622	28.927
		Std. Error Mean	3.366				
	Female	N	22				
		Mean	54.95	.12	4.67	45.82	64.50
		Std. Deviation	21.326	-.898	4.189	11.266	28.729
		Std. Error Mean	4.547				
E12: conversation score 0-100- FR	Male	N	57				
		Mean	40.89	.11	3.15	35.10	47.13
		Std. Deviation	23.516	-.281	2.013	19.158	27.246
		Std. Error Mean	3.115				
	Female	N	22				
		Mean	69.55	-.01	4.09	61.56	78.00
		Std. Deviation	18.892	-.666	2.586	13.249	23.189
		Std. Error Mean	4.028				
E19: conversation score 0-100 - RC	Male	N	57				
		Mean	62.12	.09	3.08	56.55	68.46
		Std. Deviation	23.684	-.158	1.648	20.117	26.671
		Std. Error Mean	3.137				
	Female	N	22				
		Mean	66.00	.39	6.53	52.91	78.56
		Std. Deviation	30.908	-1.495	4.555	19.679	37.136
		Std. Error Mean	6.590				
E23: conversation score 1-100 - VV	Male	N	57				
		Mean	56.58	.05	2.80	51.15	62.21
		Std. Deviation	21.535	-.307	2.311	16.725	25.948
		Std. Error Mean	2.852				
	Female	N	22				
		Mean	78.14	-.13	3.33	71.44	84.33
		Std. Deviation	15.554	-.579	2.016	10.621	18.674
		Std. Error Mean	3.316				
E41: conversation score 1-100 - RG	Male	N	57				
		Mean	46.98	-.04	3.32	39.88	53.35
		Std. Deviation	26.516	-.426	2.373	20.994	30.756
		Std. Error Mean	3.512				
	Female	N	22				
		Mean	48.00	-.06	5.07	37.96	57.86
		Std. Deviation	24.174	-.849	3.663	15.959	30.276
		Std. Error Mean	5.154				

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples