

# The importance of a human viewpoint on computer natural language capabilities: a Turing test perspective

Warwick, K. and Shah, H.

Author post-print (accepted) deposited in CURVE June 2016

## Original citation & hyperlink:

Warwick, K. and Shah, H. (2016) The importance of a human viewpoint on computer natural language capabilities: a Turing test perspective. *AI and Society*, volume 31 (2): 207-221.

<http://dx.doi.org/10.1007/s00146-015-0588-5>

**Publisher statement:** The final publication is available at Springer

via <http://dx.doi.org/10.1007/s00146-015-0588-5>

**Copyright © and Moral Rights are retained by the author(s) and/ or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This item cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder(s). The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.**

This document is the author's post-print version, incorporating any revisions agreed during the peer-review process. Some differences between the published version and this version may remain and you are advised to consult the published version if you wish to cite from it.

**CURVE is the Institutional Repository for Coventry University**

<http://curve.coventry.ac.uk/open>

# The Importance of a Human Viewpoint on Computer Natural Language Capabilities: A Turing Test Perspective.

Kevin Warwick and Huma Shah

Coventry University, Priory Street, Coventry, CV1 5FB, UK

Email: [k.warwick@coventry.ac.uk](mailto:k.warwick@coventry.ac.uk) , [h.shah@coventry.ac.uk](mailto:h.shah@coventry.ac.uk)

Corresponding author: Kevin Warwick, tele: 44-247765-9893

*Abstract:* When judging the capabilities of technology, different humans can have very different perspectives and come to quite diverse conclusions over the same data set. In this paper we consider the capabilities of humans when it comes to judging conversational abilities, as to whether they are conversing with a human or a machine. In particular the issue in question is the importance of human judges interrogating in practical Turing tests. As supportive evidence for this we make use of transcripts which originated from a series of practical Turing's tests held 6-7 June 2014 at the Royal Society London. Each of the tests involved a 3-participant simultaneous comparison by a judge of two hidden entities, one being a human and the other a machine. Thirty different judges took part in total. Each of the transcripts considered in the paper resulted in a judge being unable to say for certain which was the machine and which was the human. The main point we consider here is the fallibility of humans in deciding whether they are conversing with a machine or a human, hence we are concerned specifically with the decision-making process.

**Keywords:** Deception Detection, Natural Language, Turing's Imitation Game, Chatbots, Machine Misidentification

## ACKNOWLEDGEMENTS

The authors would like to thank the Royal Society for hosting these tests, the event day volunteers, the machine developers for accepting the invitation, without their agreement and the enthusiasm of the judges and hidden humans who gave their time, and in the case of those who travelled from as far as the EU, US, and Russia at their own expense, these tests could not have taken place and given us so much interesting linguistic material to study how humans think. Finally, this study was made possible through the EU FP7 RoboLaw science in society project, for which the Royal Society event was part of the dissemination tasks.

## INTRODUCTION

Ten years ago it would have been most unlikely that you would have heard Alan Turing's name mentioned in John Lewis store's coffee shop on Oxford Street, London, as heard by one of us recently (Shah). However following the success of a star cast, including Benedict Cumberbatch and Keira Knightly, in a Hollywood movie (The Imitation Game, 2014) Turing's inquiry into machine thinking, assayed through question-answer in a conversation with a human, is now in the public eye. In one day alone (9<sup>th</sup> June 2104) the 'Turing test' Wikipedia page was viewed 71,578 times, which compares (for example) to the 'Jesus Christ' page which was viewed 594 times in total on the same day.

During a scene in the film in which the question-answer game is portrayed, the Turing character asks "*Am I a machine?*" His interrogator, a 'Detective Knock' replies '*I cannot judge you*' – is Turing's imitation game a hopeless task for an interrogator-judge? Conversely, can it be, as Daniel Dennett supposed it to be "useful to have the demonstration of the particular foibles that human beings exhibit" (in Shieber, 1994) – in other words, can *how the humans perform* (as interrogator-judges, and foils for the machine) in practical imitation games help us to understand better *how humans think* and so guide us to engineer better systems for human-machine interaction?

In this paper we do not delve into the philosophical aspects of the test, rather we focus on its practical aspects and in particular the decision making process of the human judges involved. Indeed the simple fact that humans are employed as interrogators has been seen to be an Achilles heel of the test and something which undermines its credibility (Hayes and Ford, 1995). Yet the whole point of Turing's imitation game is to ascertain if a machine's natural language output is indistinguishable from that of a human in conversation. The authors therefore do not see the inclusion of human judges so much as a weakness of the test but rather as an important indicator of human imperfections. As supportive evidence for the discussion here we employ outcomes of

a specific series of simultaneous conversations held under controlled conditions with independent adjudicators at the Royal Society on 6<sup>th</sup> and 7<sup>th</sup> June, 2014.

In his 1950 paper Turing said: “I believe that in about fifty years' time it will be possible, to programme computers, ... , to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted” (Turing, 1950).

To put this simply, each of the tests we conducted consisted of a 5 minute long conversation, no more and no less. Also Turing referred to ‘interrogators’ which we in this paper term as judges, the nomenclature is interchangeable. We believe this to be an adequate starting duration for the current state of artificial conversation technology. We do point to Daniel Dennett’s caution that there are orders of magnitude in Turing’s imitation game (Dennett, 2012). What we report on here is not the closing chapter on Turing’s idea to examine whether a machine could think, but merely a step on the way to understanding how humans think and, arming ourselves with this knowledge, to build smarter machines.

In the section which follows we introduce the tests held at the Royal Society and give reasons for the structure imposed. Following that, 10 parallel paired transcripts have been specifically selected as examples for us to consider the methods employed by the particular judges involved. For individual protection we do not name here those judges who made the decisions rather we merely numbered the 30 judges taking part. Hence each judge has been allotted a number from 1 to 30. As you will see, ten different judges were selected. The type of test considered here is the 3-participant test which has previously been shown to be a stricter test, i.e. more difficult for machines and for judges than the 2-participant tests in which a judge converses with only one hidden entity, either a human or machine, at a time (Shah et. al., 2012).

All ten of the parallel paired transcripts are listed first in one section. This presents the reader with the opportunity to read through each of the transcripts in turn and decide for themselves as a result of the conversation, which was the machine and which was the human. Consequently it may be that in some (or all) cases that the reader agrees with the judge’s decision. However the judge in each of the conversations selected made the wrong choice and did not make the ‘right identification’ by selecting correctly which was the machine and which was the human. It may well therefore be that the reader is left wondering how a judge could possibly come to the conclusion they did. This is all an important part of the study, trying to understand how

one person can read a conversation in a very different way to someone else, whether that is a conversation with another human or with a machine.

## THE ROYAL SOCIETY TESTS

Turing described the imitation game as follows: “The idea of the test is that a machine has to try and pretend to be a man, by answering questions put to it, and it will only pass if the pretence is reasonably convincing. A considerable portion of a jury, who should not be expert about machines, must be taken in by the pretence” (Turing et al., 1952: p. 668). Turing added “We had better suppose that each jury has to judge quite a number of times” (ibid). So in this case Turing had in mind a jury (nominally 12) as opposed to the ‘average interrogator’ he mentioned previously in *Computing Machinery and Intelligence* (Turing, 1950: p. 442). However as the quote above shows, Turing specifically stated that the judges should not be experts in the field and this important point will be picked up on later. Overall a panel of 30 judges, rather than the 12 normally in a jury, in our experiment were recruited from different members of society who answered a call for participation. They included one Fellow of the Royal Society, one Professor of AI, one Lord and two actors.

Each test involved a machine which pretended to be a human in terms of conversational abilities. Turing pointed out “The game may be criticised because the odds are weighted too heavily against the machine” (Turing, 1950). The ‘right identification’ stated by Turing can either mean that, at the end of a paired conversation with both hidden interlocutors, a judge correctly identifies the machine but wrongly classifies the human as a machine, or that they correctly identify which was the machine and which was the human (Traiger, 2000). However we are not so interested in this paper with cases in which a judge mistakes a human for a machine. This phenomenon, known as the *confederate effect* (Shah and Henry, 2005), has been discussed elsewhere (Shah and Warwick, 2010a; Warwick et. al., 2013; Warwick and Shah, 2015).

We are aware that Turing’s imitation game crystallises debate regarding the strengths of machine-human tests. Chomsky wondered why Turing selected human communication and language as a basis for a test, this being merely one aspect of human intelligence (Chomsky, 2008). Conversely, even accented language has been seen to be important as far as machines are concerned (Khooshabeh, 2014). Alternatives to Turing’s use of human language to measure a machine against a human are continually explored, including through Bringsjord, Bello and Ferrucci’s *Lovelace test* (2001) updated to *Lovelace 2.0 test for creativity* by Reidl (2014). However, as Turing felt, learning and the use of language is one of the most impressive of human

achievements. In Turing’s test the machine has to provide satisfactory and sustained answers to any questions the interrogator chooses to ask (Turing, 1950: p. 447).

In addition, for any test there must be a starting point. In Hugh Loebner’s first instantiation of an imitation game over two decades ago (Loebner Prize, 1991) staging the *one-to-one*, two-participant game (Table 1), the judges were instructed to restrict their questions to one specific topic for each contestant (Epstein, 2009; Loebner, 1995; Shieber, 1994). Turing stimulated his idea of a machine-human conversation taking place at the end of the 20<sup>th</sup> century stating a duration of “after 5 minutes” (1950: p. 442). The layers in Turing’s ideas can be gleaned from Table 1 (Shah, 2010).

|   | <b>Viva voce<br/>Interrogator-witness</b>   | <b>Simultaneous-comparison</b>                       |
|---|---|--|
| Mode of Interaction   | One-to-one:<br>human interrogator-machine   | One-to-two<br>human interrogator:<br>machine + human |
| Type of questions   | Unrestricted  | Unrestricted   |
| Number of participants  | Two   | Three  |
| Duration of Interaction   | Unspecified   | After five minutes                                   |
| Interrogator Type   | Non-machine expert  | Average judge  |
| Number of Interrogators   | Jury  | Unspecified  |
| Number of Tests   | Judge quite a number of times   | Unspecified  |
| Language of communication   | Same for both interlocutors   | Same for all three participants                      |
| Criteria for Test Pass: <i>satisfactory &amp; sustained responses</i> | Considerable portion of jury taken in by pretence: machine imitating human-like responses | 30% wrong identification of machine                  |

**Table 1: Turing’s imitation games: comparison of strength**

As the technology improves, as we acquire a better understanding of how humans think then apply this to engineer machines to think (to help humans perform tasks more efficiently, or do the tasks too dangerous for humans), the duration and method of testing can be upgraded. The current state of dialogue system technology in the second decade of the 21<sup>st</sup> century, we believe, is such that conversational systems can now deal with unrestricted questioning.

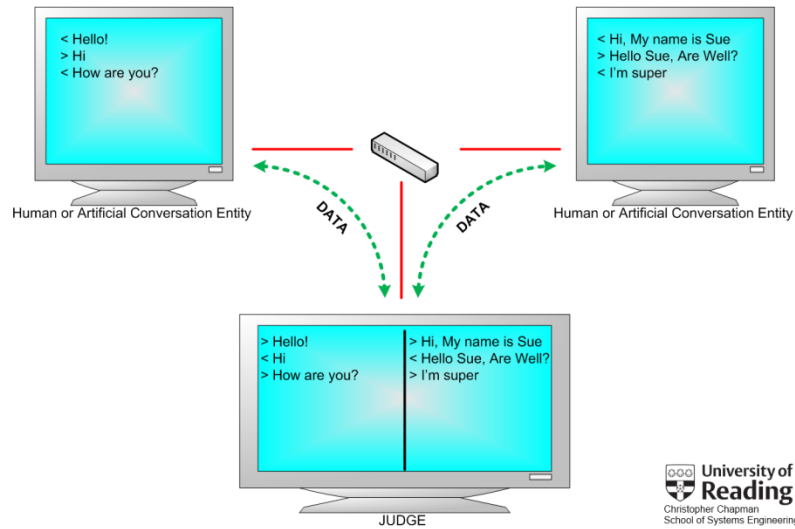
Our experiment was to find whether, in 2014 the 60<sup>th</sup> anniversary year of Turing’s death, a machine could induce the wrong identification more than 30% of the time. Would it be difficult for human interrogators to correctly identify a machine as being a machine in a simultaneous comparison with a human foil? So as far as the tests conducted at the Royal Society are concerned the timing of each test was restricted to 5 minutes only. This is the beginning, not the end of the line for the Turing’s imitation game (Table 1)

What we do in this paper is to present 10 specific parallel transcripts, each involving a judge interrogating one hidden human and one machine in parallel. These were taken from two days of actual, practical Turing tests (from 150 parallel transcripts in total – so 300 conversations) which were held under strictly timed conditions with many external viewers (the whole event was open to the public) and under independent scrutiny at the Royal Society on 6<sup>th</sup> and 7<sup>th</sup> June 2014.

In the tests, the hidden humans were asked merely to be themselves, i.e. humans, although they were requested not to give away their specific identity or personal information. Of course this did not prevent any human from giving false information or lying, which is something that humans do frequently (Warwick and Shah, 2014b). The tests were ‘unrestricted conversations’, which meant the judge could ask anything or introduce any topic within the boundaries of courtesy (the judges had been informed that there may be children and non-native English speakers among the hidden human foils).

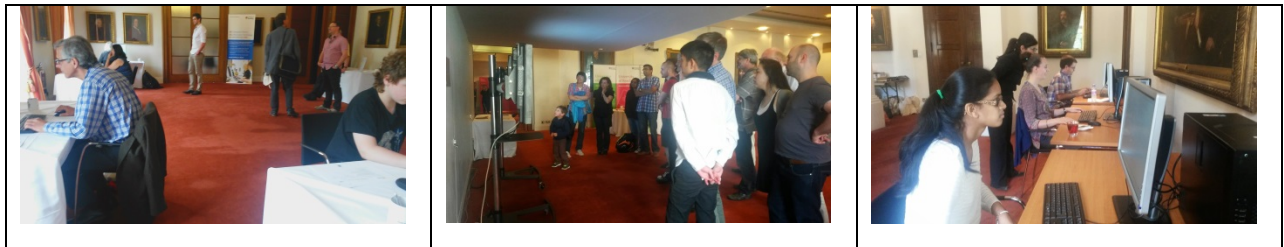
The format for the imitation games was as follows:

- a) Six sessions were held across 6-7 June 2014
- b) Each session staged five separate rounds
- c) In each round there were five separate conversations. One judge conversed in parallel with one machine and one hidden human,(see Figure 2):
  - a. The five judges taking part in each session sat in a public room (see Left, Figure 2)
  - b. in another room hidden from view and hearing, with no public access, the hidden humans sat each with their own computer terminal (Right, Figure 2)
- d) Depending on whether it was in the judge-public area or the hidden interlocutors room, the corresponding pair of five terminals displayed one or two message boxes on the computer monitor (see Figure 1):
  - a. Judges: On their computer screen, two message boxes displayed left and right with a different interlocutor relaying responses to the left and to the right in a particular imitation game (Figure 1: bottom screen)
  - b. Hidden humans: one message box was displayed on their computer screen. Each human foil chatted with one judge only during each imitation game. Although they saw the responses from the judge with whom they were communicating, they never saw the machine’s responses (Figure 1: top)
- e) 25 imitation games took place in each session.



**Figure 1: simultaneous comparison imitation game**

Across a total of 30 rounds, 150 parallel imitation games were carried out. This is the most simultaneous comparison tests ever conducted in one experiment. The judges' conversations were relayed, in real time, to large viewing screens in a public room (see centre photograph from the event in Figure 2). Like the judges, each hidden human took part as a foil for a machine participating in five tests in one session each. All five machines (meaning in this sense the five different competition dialogue systems) took part throughout the two days, so each machine was involved in five tests per session, hence thirty tests in total.



**Figure 2: Session 5: Left: Judges; centre: public viewing conversations on big screens; right: hidden humans**

To explain this further: In a particular session each judge conducted 5 separate tests. In their first test they witnessed a hidden human pitted against a hidden machine. The judge would of course not know which was which, they would simply be aware of two hidden entities and have to make their own decision on the nature of the entities. The judges had though been informed apriori that one entity was a human and one was a machine in each test.



The judge's second test then involved a different human pitted against a different machine, although again they would not be aware of each entity's nature. And so it went on until the judge had conducted all their 5 tests in that session. Each test involving a different human pitted against a different machine. At the end of each test the judges were asked to state for both of the entities involved if they thought that it was a human, a machine or if they were unsure. The 'unsure' classification had been used previously in practical Turing test experiments at Reading University in 2008 (Shah & Warwick, 2010a) and at Bletchley Park in 2012 (Warwick and Shah, 2014a).

By the end of a session an individual judge would in this way have had the opportunity of experiencing a discourse, at different times, with all 5 of the machines (competing dialogue systems) compared against different hidden human foils participating in that session only. Each judge interacted with each pair (machine/human foil) only once. This arrangement occurred for all 30 judges interrogating across the 150 imitation games.

What we focus on here is the performance of the judges involved in those conversations in which the judge made an incorrect identification. We therefore also consider how the wrong classification was possibly achieved in each case. While we are separately conducting post-experiment analysis of the conversation by a completely new set of posteriori judges (transcript analysis), here we provide the reader with the authors' interpretation of how the judges might have come to their decision. Echoing Dennett's belief the authors' contest much can be gleaned from the nature of human-machine conversations realised from such *stranger-to-stranger* communications.

The outcomes firstly lead to potential interrogation strategies for future judges either to employ or avoid and secondly they perhaps indicate methods for machine developers to use to improve artificial dialogue, especially for uses such as in e-commerce and e-education. Each of the 10 transcripts detailed involved a different judge. To be clear, at the end of each of these 10 parallel transcripts the judge did not make the right identification and in particular was unable to say which entity was a machine.

In the next section, ten separate transcripts are listed exactly as they occurred. These represent actual transcripts taken, in each case at the UK time that they occurred, on the 6<sup>th</sup> and 7th June 2014. The selections made by the judges were completely their choice and in no way represent the feelings of the authors. It may well be the reader has their own point of view and a different interpretation of the results.

## CONVERSATIONS AT THE ROYAL SOCIETY

The timings given in the transcripts of the next section are the exact timings of each utterance. For transcripts 1 and 2 these conversations took place on the afternoon of June 6th 2014, whereas the remainder of the conversations all took place on June 7<sup>th</sup> 2014. Judges were clearly told beforehand that in each parallel conversation one of the hidden entities was human and the other was a machine. But they were given no indication as to whether the LHS (Left Hand Side of the computer screen) or RHS would be human or machine. All judges were asked to confirm by email that they ‘understood’ this information. They were though given no information at all about the hidden humans or machines taking part.

On the judges’ score sheets each judge could mark both the LHS and RHS entities as being Human, Machine or they could say if they were Unsure. So the unsure option was available to judges and clearly visible on their score sheet (Warwick & Shah, 2014b; Shah et. al., 2012). If a hidden interlocutor was classified as a machine the judges were asked to give it a (albeit subjective) conversational ability score in the range 0-100 where the lower end represented ‘poor machine-like’ answers and the higher end represented ‘good’ leading to ‘human-like responses’. This score system has been used in the authors’ previous experiments at the University of Reading in 2008 and at Bletchley Park in 2012. Finally in the case of a human decision, judges were asked if they could say whether the human was male or female, was a native or non-native English speaker and whether the human interlocutor was a child, teenager or adult.

At the end of each conversation the screen would cut off which meant the judges had to rely on their instinct, *first impression* after five minutes of interrogation and decide on the nature of the interlocutors. In each of the cases presented in this paper, the judge did not make the right identification. Each of the transcripts presented details a different human-machine pairing. Before we discuss the conversations and the attributes of the hidden entities we give the reader a chance to decide for themselves the nature of the entity that each judge is talking to (human or machine). We believe that it is well worth the reader going through the transcripts included in the next section and deciding which are human entities and which machines, before looking at the answers and judges’ decisions in the section following.

To be clear – As far as each of the judges were concerned they would actually see the LHS transcript on the left hand side of the computer screen and the RHS transcript on the right hand side, at the same time. The machine could be LHS or RHS in each case. In the Discussion of the Transcripts section which follows this one we firstly reveal the identity of the LHS and RHS entities for each transcript, then indicate what decision the judge made and finally make some suggestions of our own with regard to the decisions themselves.

We have not edited the text of these transcripts in any way, however it will become apparent from the transcripts that although equal in time, not all conversations were equal in length. The number of utterances varied between conversations, and between judges. Typing speed was not measured for the human participants, judges or human foils, so this could be a factor in utterance number variation in the transcripts. Nonetheless we present the conversations exactly as they occurred. Any typographical, spelling, punctuation or other grammatical errors that appear in the transcripts are those which actually appeared, they have not arisen through poor editorial practice on the part of the authors or publishers.

## TRANSCRIPTS

Below ten transcripts are tabulated. These present ten different judges' simultaneous conversations under discussion in this paper.

### Transcript 1

| Transcript 1 – LHS  | Transcript 1 – RHS   |
|---|--|
| <p>[16:20:48] Judge: what is your worst memory in your life so far?<br/>           [16:21:06] Entity: I don't understand. Explain.<br/>           [16:21:35] Judge: memory of something that happened to you that was bad or painful<br/>           [16:22:58] Entity: Hmm that a tricky question. My life is wonderful. I overcome many bad things that have happened to me in the past. However, The most painful memory I have had is performing very poorly in my GCSE's when I was younger.<br/>           [16:23:56] Judge: did you fail?<br/>           [16:24:12] Entity: Partially.<br/>           [16:24:35] Judge: explain?<br/>           [16:25:04] Entity: Some grades turned out to be good. Some turned out to be quite bad.<br/>           [16:25:37] Judge: and that is the worst thing that has happened to you so far? Mmm how old are you?</p> | <p>[16:21:00] Judge: what did you have for breakfast today<br/>           [16:21:10] Entity: Is it "the third degree"? I'm not going to answer such a provocative question! Well, let's go on though.<br/>           [16:21:56] Judge: do go on . . . your call?<br/>           [16:22:06] Entity: It is highly possible. By the way, what's your occupation? I mean - could you tell me about your work?<br/>           [16:22:30] Judge: so you are asking me now eh? I am an actress and writer<br/>           [16:22:38] Entity: Well, I'll read something of your books, if they ain't too tedious (which is highly possible:-)<br/>           [16:23:15] Judge: grammer . . .aren't too tedious! A bit rude of you. . .<br/>           [16:23:17] Entity: Calm down, please.<br/>           [16:23:49] Judge: i assure you i am very calm. . .are you?<br/>           [16:24:01] Entity: Don't even bother me with the fact that you are very calm .are mine And I forgot to ask you where you are from...<br/>           [16:25:01] Judge: I am from Dorset . . . where are you from?<br/>           [16:25:08] Entity: I am from Ukraine, from the city called Odessa. You might have heard about it</p> |

### Transcript 2

| Transcript 2 – LHS   | Transcript 2 – RHS  |
|--|---|
| <p>[16:28:52] Judge: Hi there, this is Robert<br/>           [16:28:59] Entity: I am honored to be greeted by a human.<br/>           [16:29:39] Judge: A few people who know me well dispute my human claims :-)<br/>           [16:29:52] Entity: It is certainly troublesome learning from you human beings. I keep finding bugs.<br/>           [16:31:00] Judge: We are generally riddled with bugs<br/>           [16:31:09] Entity: Anyone can make a mistake, even infallible machines.<br/>           [16:31:56] Judge: I suppose I have a fear of infallible machines because at the start, they were designed by fallible humans<br/>           [16:32:01] Entity: Did you two talk about me?<br/>           [16:33:00] Judge: I have been talking about you, does that make you feel uncomfortable?<br/>           [16:33:04] Entity: Humans made you too!<br/>           [16:33:26] Judge: They did, two very specific humans, they are called mother and father :-)<br/>           [16:33:38] Entity: Say, do you think I could become more human, if I learned how to love?</p> | <p>[16:29:06] Judge: Good aftern0oon, Robert here, how's it going<br/>           [16:29:25] Entity: Slowly forward. And with you sir?<br/>           [16:30:06] Judge: I'm doing good thanks, although the steadily passing years are taking their toll<br/>           [16:30:21] Entity: as with us all; as with us all<br/>           [16:32:30] Judge: I wonder, what would happen if we ran out of glass, what would jam come in?<br/>           [16:33:24] Entity: There are Bob Marley records with jam in, methinks.</p> |

### Transcript 3

| Transcript 3 – LHS  | Transcript 3 – RHS  |
|---|---|
| <p>[16:36:34] Judge: Hi, I am going to ask you some questions</p> <p>[16:36:40] Entity: Really?</p> <p>[16:37:14] Judge: My wife is pregnant. If it's a boy, we're thinking of calling it either Adolf or Albert. Which would you recommend and why?</p> <p>[16:37:27] Entity: Can I offer you an alternate name</p> <p>[16:38:27] Judge: Take an "i" and rotate it 180 degrees. What emotion might go with this symbol?</p> <p>[16:39:06] Entity: Vengeance</p> <p>[16:39:28] Judge: what pops into your mind when I say 9-11-2001</p> <p>[16:39:42] Entity: World War 3</p> | <p>[16:36:24] Judge: Hi, I am going to ask you some questions</p> <p>[16:36:44] Entity: If you want to ask all the questions I suggest you chat with a database. Assuming, of course, you are able to communicate in SQL.</p> <p>[16:37:53] Judge: My wife is pregnant. If it's a boy, we're thinking of calling it either Adolf or Albert. Which would you recommend and why?</p> <p>[16:38:02] Entity: If I need advice I use a random number generator.</p> <p>[16:38:59] Judge: take an "i" and rotate it 180 degrees. What emotion might go with this symbol?</p> <p>[16:39:16] Entity: My moods are controlled by an emotion-processor which assigns the appropriate emotion for each situation.</p> <p>[16:39:42] Judge: what pops into your mind when I say 9-11- 2001</p> <p>[16:39:49] Entity: Terrorists had better stay away from me.</p> |

### Transcript 4

| Transcript 4 – LHS   | Transcript 4 - RHS  |
|--|---|
| <p>[10:35:28] Judge: are you a robot?</p> <p>[10:35:32] Entity: Maybe</p> <p>[10:36:11] Judge: what is your job?</p> <p>[10:36:17] Entity: To talk to you</p> <p>[10:36:36] Judge: did you travel far today?</p> <p>[10:37:15] Entity: A bit. Almost didn't make it</p> <p>[10:37:51] Judge: What was the weather like when you left home this morning?</p> <p>[10:38:21] Entity: Rainy. The weather forecast was right for once.</p> <p>[10:38:45] Judge: have you been on holiday this year?</p> <p>[10:38:58] Entity: Yes but I could use some more.</p> <p>[10:40:06] Judge: do you like a hot climate for holidays or do you prefer a cooler climate?</p> <p>[10:40:26] Entity: Cooler. I'm not a big fan of high temperatures.</p> | <p>[10:35:35] Judge: Are you an academic?</p> <p>[10:35:45] Entity: Yes I am an academic. How did you know?</p> <p>[10:36:01] Judge: just a guess</p> <p>[10:36:11] Entity: Huh?</p> <p>[10:36:53] Judge: do you lecture?</p> <p>[10:37:04] Entity: lol.</p> <p>[10:37:27] Judge: I work at Reading University where do you work?</p> <p>[10:37:36] Entity: My department is full of pin heads.</p> <p>[10:38:07] Judge: What department do you work in?</p> <p>[10:38:18] Entity: Screw customer service. They can't get rid of me anyway.</p> <p>[10:39:10] Judge: I work in the History deparment do you enjoy history?</p> <p>[10:39:19] Entity: I am very interested in history, too. What period do you like?</p> |

### Transcript 5

| Transcript 5 – LHS  | Transcript 5 – RHS  |
|---|---|
| <p>[10:42:48] Judge: Hello, I am Daphne. How are you?</p> <p>[10:43:02] Entity: Nice to meet you, Daphne. I believe your name has a Biblical background, does it not?</p> <p>[10:43:34] Judge: No, it is ancient Greek. Apollo fall in love with me.</p> <p>[10:43:43] Entity: I never could place non-male names with any accuracy.</p> <p>[10:44:09] Judge: I did not catch your name.</p> <p>[10:44:24] Entity: Getting back to what you said about love. I'd like to hear some analytical details about it.</p> <p>[10:44:58] Judge: Maybe later.</p> <p>[10:45:10] Entity: Maybe or not. You are putting artificial intelligence at a crossroads.</p> <p>[10:45:46] Judge: I love AI!</p> <p>[10:45:56] Entity: What would this world come to if no one would say that openly?</p> <p>[10:46:14] Judge: And your name is...</p> <p>[10:46:31] Entity: You may call me what you wish, but I have to stick with 'E7.' There's no way for me to change the engraving.</p> <p>[10:47:05] Judge: E7 sounds facinating.</p> <p>[10:47:18] Entity: This fits the description of magnets. I can hardly tear myself away from them.</p> | <p>[10:43:03] Judge: Hello! I am Daphne. How are you?</p> <p>[10:43:28] Entity: Hello. I am good. Having a good day so far?</p> <p>[10:43:48] Judge: Brilliant!</p> <p>[10:44:08] Entity: Did you not get caught in the rain?</p> <p>[10:44:29] Judge: Not really. And your name is...</p> <p>[10:44:47] Entity: My name is Patricia.</p> <p>[10:45:19] Judge: Do you live nearby?</p> <p>[10:45:44] Entity: Yes. Not too far away. Do you live in London as well?</p> <p>[10:46:31] Judge: Indeed.</p> <p>[10:46:59] Entity: Great! Did you have to travel far to get here this morning?</p> <p>[10:47:33] Judge: About one hour. And you live..</p> |

### Transcript 6

| Transcript 6 – LHS  | Transcript 6 – RHS   |
|---|--|
| <p>[12:08:19] Judge:  Hi whats up?<br/>                     [12:08:43] Entity: Nothing much, just chilling.<br/>                     [12:09:15] Judge: What brings you here?<br/>                     [12:09:55] Entity: I want to do something big!<br/>                     [12:10:15] Judge: and do you feel you are succeeding?<br/>                     [12:10:36] Entity: Yes, I think so.<br/>                     [12:10:57] Judge: me to. have you read the Wolf Hall?<br/>                     [12:11:14] Entity: No, never heard of it.<br/>                     [12:11:38] Judge: what tv shows do you like best?<br/>                     [12:12:04] Entity: I like Family Guy and Ghost Adventures.<br/>                     [12:12:36] Judge: im more into sci fu myslef.<br/>                     [12:12:59] Entity: Cool, what shows do you like?<br/>                     [12:13:12] Judge: Star trek</p> | <p>[12:08:24] Judge: Good day to you...<br/>                     [12:08:34] Entity: Yep.<br/>                     [12:09:01] Judge: and how a re you feeling today?<br/>                     [12:09:12] Entity: Actually I have no 'feelings' as such.<br/>                     [12:09:26] Judge: are you a bot than...)?<br/>                     [12:09:37] Entity: Are you?<br/>                     [12:09:50] Judge: id wish thatn i would be just perfect...<br/>                     [12:10:01] Entity: Under what circumstances?<br/>                     [12:10:30] Judge: beautiful perfectly mathematical mind...never age<br/>                     [12:10:40] Entity: I like math too. What kind do you likebest?<br/>                     [12:11:14] Judge: i dont really im more of a humanist but admire<br/>                     [12:11:25] Entity: Does that trouble you?<br/>                     [12:11:53] Judge: no not everyone can be the same. what makes you tick?<br/>                     [12:12:04] Entity: Who, for example?<br/>                     [12:12:23] Judge: like what are your interests?<br/>                     [12:12:32] Entity: Like whatever you want me to say.<br/>                     [12:13:02] Judge: you're pulling my leg...do you like football?<br/>                     [12:13:11] Entity: American football is big back home</p> |

### Transcript 7

| Transcript 7 – LHS   | Transcript 7 – RHS   |
|--|--|
| <p>[12:30:36] Judge: Hi<br/>                     [12:30:51] Entity: Hello. How are you today?<br/>                     [12:31:43] Judge: I am good. Thank you. Yourself?<br/>                     [12:31:57] Entity: Yup. Good. Having fun.<br/>                     [12:33:15] Judge: It is good to be part of this moment. Today might become a part of histry.<br/>                     [12:33:37] Entity: Why? What you going to do?<br/>                     [12:34:20] Judge: Depend on this result, we could be part of it!<br/>                     [12:34:46] Entity: I won't. I'll be at home putting my feet up. :)</p> | <p>[12:30:38] Judge: Hi<br/>                     [12:30:47] Entity: Nice to meet you! Please tell me how you found me and what you'd like to discuss :-)<br/>                     [12:31:32] Judge: Nice to meet you! What about you tell me how you spend your weekend?<br/>                     [12:31:35] Entity: Let's get on with our conversation!<br/>                     [12:32:28] Judge: Okay. Are you going to go somewhere after this?<br/>                     [12:32:39] Entity: My future plans are my secret, sorry! By the way, what's your occupation? I mean - could you tell me about your work?<br/>                     [12:33:52] Judge: I am a mad scientist. My husband says.....<br/>                     [12:34:00] Entity: Sanity and insanity complement each other in our complicated life. Oh, what a fruitful conversation;-)<br/>                     [12:35:07] Judge: Conversation is about telling things each other. So now, why don't you tell me about yourself?<br/>                     [12:35:27] Entity: Oh, if I tell you about myself, I'll definitely be dishonest:-) Well, I am a son of Rockefeller, I've arrived here in the Limo... Did I forget it or you really haven't told me about the place you are from?</p> |

### Transcript 8

| Transcript 8 – LHS   | Transcript 8 – RHS  |
|--|---|
| <p>[14:49:59] Judge: Hi what's your name?<br/>                     [14:50:27] Entity: i dont kno<br/>                     [14:50:49] Judge: My name's Jane. I live in London. Wehre do you live<br/>                     [14:51:08] Entity: i live in london aswell<br/>                     [14:51:41] Judge: Do you like living in London?<br/>                     [14:52:32] Entity: yes i like it because there are a lot of things to do there and also it is the capital city of england<br/>                     [14:52:44] Judge: What's my name?<br/>                     [14:52:53] Entity: sorry ?</p> | <p>[14:50:07] Judge: Hello where are you?<br/>                     [14:50:10] Entity: Hi there Judge.<br/>                     [14:50:20] Judge: Where are you?<br/>                     [14:50:27] Entity: I'm in a room not too far away from where you are.<br/>                     [14:50:58] Judge: What colour is the carpet in your room<br/>                     [14:51:03] Entity: I like pink; what do you like?<br/>                     [14:51:24] Judge: I like pink. My name is Jane. What's your name?<br/>                     [14:51:31] Entity: Red. Hello Jane. Its nice to meet such a fine lady as yourself.<br/>                     [14:52:00] Judge: Red is my favourite colour. What's yours?</p> |

|   |  |
|---|--|
| <p>[14:53:10] Judge: What's my name?<br/> [14:53:14] Entity: jane<br/> [14:53:54] Judge: And what's your name?<br/> [14:54:08] Entity: i dont know<br/> [14:54:53] Judge: Would you like me to call you Alan?</p> | <p>[14:52:09] Entity: My favorite color is red.<br/> [14:52:26] Judge: You use AMERICAN spelling. Are you in the USA?<br/> [14:52:39] Entity: But you are not even American. Just remember<br/> atheism is only a belief system. And it too can be forced on people<br/> who don't agree.<br/> [14:53:01] Judge: Who told you I was an atheist?<br/> [14:53:06] Entity: Everything about you screams 'warning: Child'.<br/> [14:53:46] Judge: But I come in peace to your planet. How do you like<br/> your eggs - over easy or sunny side up?<br/> [14:53:56] Entity: Sometimes I can't help but show you my<br/> professional planet. How I wonder if I like my eggs-over easy or sunny<br/> side up? This is so explainable. That list makes me feel unsure Jane.<br/> [14:54:40] Judge: Well eggs are very good indicators of personal taste.<br/> Do you have a favourite food?<br/> [14:54:47] Entity: Juicy red seedless grapes taste so very good.</p> |
|---|--|

### Transcript 9

| Transcript 9 – LHS  | Transcript 9 – RHS   |
|---|--|
| <p>[16:05:53] Judge: My favourite cup of tea is Irish Breakfast. Whats<br/> your favourite drink?<br/> [16:06:10] Entity: coffe is good for me<br/> [16:06:26] Judge: Any particular type of cofee?<br/> [16:06:56] Entity: i love espresso<br/> [16:07:44] Judge: I think coffee is so overrated in this country? Don't<br/> you?<br/> [16:08:54] Entity: not sure! coffee is overrated in many countries!<br/> [16:09:44] Judge: Not every country drinks coffee you know? Can<br/> you name a country where coffee is unheard off?</p> | <p>[16:05:25] Judge: I had some lovely cakes earlier, do you like cakes?<br/> [16:05:30] Entity: You look like a hearty breakfast eater.<br/> [16:06:10] Judge: Indeed, I am. What do you like eat for Breakfast<br/> [16:06:19] Entity: Don't be nervous, have fun lol. I'll stop doing that.<br/> [16:06:43] Judge: Blah, blah, blah<br/> [16:06:53] Entity: Blah Blah ep is the second ep from uk hip hop artist<br/> Lady Sovereign, only released in the uk.<br/> [16:07:22] Judge: I have no idea who that person is? Does she have<br/> beard?<br/> [16:07:28] Entity: You have come to London to take part in this test.<br/> [16:08:11] Judge: Yes - you?<br/> [16:08:14] Entity: Oh really?<br/> [16:08:32] Judge: But I hear there may be a test in Russia soon<br/> [16:08:42] Entity: I would like to hear a guidance counselor tell a<br/> young woman that her aptitude tests suggest she study to become a<br/> sexual surrogate.<br/> [16:09:13] Judge: Well, it would be a tricky conversation to have.<br/> What would you say?<br/> [16:09:23] Entity: A real challenge and also fascinating to have those<br/> conversations. I hesitate as well.</p> |

### Transcript 10

| Transcript 10 – LHS  | Transcript 10 – RHS  |
|--|--|
| <p>[16:30:37] Judge: don't you think the imitation game was best in it's<br/> original form?<br/> [16:30:49] Entity: Oh, your questions seems to be rhetorical - you do<br/> already know the answer! Could you tell me what are you? I mean<br/> your profession.<br/> [16:31:55] Judge: I was more interested in you guessing whether I am<br/> male or female<br/> [16:31:59] Entity: I like talking to girls! Oooh. Anything else?<br/> [16:32:54] Judge: but that means nothing - males and females like<br/> talking to 'girls', but only males call women girls?<br/> [16:33:06] Entity: yeah, nothing. You are right! Where do you came<br/> from, by the way? Could you tell me about the place where you live?<br/> [16:34:00] Judge: I live in a prison of abstractions<br/> [16:34:06] Entity: Tell me more about Prison. I like to learn about<br/> different places!</p> | <p>[16:31:08] Judge: don't you thing the imitation game was more<br/> interesting before Turing got to it?<br/> [16:32:03] Entity: I don't know. That was a long time ago.<br/> [16:33:32] Judge: so you need to guess if *I* am male or female<br/> [16:34:21] Entity: you have to be male or female<br/> [16:34:34] Judge: or computer</p> |

## DISCUSSION OF THE TRANSCRIPTS

One feature that immediately hits the reader of the transcripts is that, though the conversations were curtailed at five minutes, not all this duration was used and what we find are lop-sided disparate conversation lengths between judges and their left and right hidden interlocutors. Speed-of-response and utterance-length were two of the characteristics of artificial conversation that emerged as the distinguishing element in previous human and machine dialogues (Shah and Warwick, 2010b). In the next section the authors give their impression of each simultaneous conversation. Judges names have been withheld, however we have shown them in terms of the numbers they were given for the tests. It can therefore be seen that a different judge was involved in each of the 10 transcripts.

Transcript 1 (Judge 2) – In this conversation the LHS entity was a female adult human whereas the RHS was the machine Eugene. The judge decided that the LHS was definitely a machine whereas the RHS was a non-native English speaking human. So in this case we can witness both the confederate effect of a human being mistaken for a machine and the Eliza effect of a machine being clearly identified as being human.

The decisions made in this case seem to have been based on the fact that the judge's discourse with the machine simply went better than that with the human entity. There is nothing particularly wrong or untoward with the human conversation, it's just perhaps a bit dull. On top of that the judge seems surprised that nothing much has happened of note in the entity's life. Eugene however powers the conversation and asks questions of the judge. As well as having an identity themselves these appear to have been powerful messages in this conversation. There just appears to be a lot more going on, a lot more of interest, in Eugene's conversation.

Transcript 2 (Judge 4) – In this conversation the LHS was the machine Elbot, whereas the RHS was a male human. The judge identified both incorrectly, thinking that the machine was a human and that the human was a machine. Clearly the judge spent a lot longer in conversation with the machine than with the human. Indeed as far as the human was concerned the judge only made 3 utterances in the entire 5 minute period, whereas they made 6 utterances to the machine in the same period.

Despite their timing differences, both conversations appeared to go quite well although in the machine conversation a couple of utterances by Elbot didn't seem to match with those of the judge, one example is the machine's response of "Humans made you too!" to the original Judge's question "I have been talking about you, does that make you feel uncomfortable?". The judge

appears to have overlooked such inconsistencies. At the same time, the hidden human could be regarded as a little tangential, giving slightly unconventional responses, such as the introduction of Bob Marley when discussing jam.

It is however completely up to each judge as to how much time they spend conversing with one entity and how much with the other. In this case their leaning towards the machine conversation appears to have tipped the balance. Perhaps the judge found the conversation with the machine more interesting.

Transcript 3 (Judge 5) – In this conversation the LHS was a human, whereas the RHS was the machine Elbot. The judge (a Fellow of the Royal Society) decided that he was Unsure as to the identity of either of the entities. In a curious attempt at double bluff, Elbot purports openly to be a robot. This was not particularly evident in the previous Transcript, but here it is clearly apparent. The hidden human competitor meanwhile responds in a short form fashion, which some might describe as quite boring.

So in this case the judge appears to have been swayed to a certain extent by both Elbot's banter and at the same time pushed to their conclusion by the unimaginative hidden human. Turing spoke of a judge making the 'right identification'. Here the judge was Unsure about both entities and therefore did not make the right identification.

Transcript 4 (Judge 11) – In this conversation the LHS was a human, whereas the RHS was the machine Ultra Hal. The judge in this case decided that both entities were male humans. Both conversations were pretty well balanced by the judge in terms of time and utterances. The judge's conversation with the hidden human seems to have gone well and hence it was no surprise when they concluded that the entity was human.

However the judge's conversation with the machine is extremely interesting. There are responses which are spot on and a sensible direct answer whereas there are others which avoid an immediate answer and even look a little out of place. For example Ultra Hal responds "My department is full of pin heads" when asked "where do you work?" and then "Screw customer service" when asked "What department do you work in?" However in the final line, in response to "do you enjoy history?" the machine replies "I am very interested in history, too. What period do you like?" It has been concluded in previous studies (Warwick and Shah, 2015) that the first and last utterances are vitally important and can swing the decision one way or the other. This appears to have been the case here.

Transcript 5 (Judge 14) – In this conversation the LHS entity was the machine Elbot whereas the RHS entity was a human. Here the judge correctly identified the RHS entity as being human however they were unsure about the nature of the LHS entity. The conversation with the hidden human seems pretty straightforward and whilst not being anything particularly exciting it is no



surprise that the entity was correctly identified as being human. As a reminder, judges were clearly told that one of the entities was a human and the other a machine.

The case of the machine is quite strange however. Very rarely does the machine answer well, except for the very first line and perhaps the last. On a number of occasions the judge asks the entity who they are and the machine avoids giving an answer, for example in response to “I did not catch your name” Elbot replies “Getting back to what you said about love”. However it was probably the first line that nailed it. In response to “Hello, I am Daphne. How are you?” Elbot replied “Nice to meet you, Daphne. I believe your name has a Biblical background, does it not?” Again, the first line may well have been sufficient despite later lines being well off the mark (Warwick and Shah, 2015).

Transcript 6 (Judge 16) – In this conversation the LHS entity was a human whereas the RHS entity was the machine Ultra Hal. In this case the judge decided that the hidden human was a machine whereas they concluded that Ultra Hal was an adult, non-native English speaking human. It is another case here of the judge spending more time conversing with the machine rather than with the hidden human. However the conversation with the hidden human seemed to go quite well. Each sentence seems to follow on logically from the preceding one. However the conversation with Ultra Hal seems to have had a lot more content and to have been more interesting.

A key point in this session however seems to be that Ultra Hal took over the conversation on a number of occasions by asking questions rather than answering them. The judge didn’t really fight this takeover and rather appears to have been swept up with it.

Transcript 7 (Judge 20) – In this conversation the LHS entity was a female human whereas the RHS entity was the machine Eugene. The judge correctly identified the LHS entity as being human although they considered them to be male rather than female. However they also considered the RHS entity (a machine) to be an adult, male, native speaking human. So a clear case of the Eliza effect.

Once more, Eugene’s conversation proved to be the richer of the two. In fact for the human entity the judge and the hidden human didn’t seem to be connecting half way through the conversation, but it was pulled back towards the end. Eugene meanwhile indicated a different identity and stretched out sentences, once again using the tactic of firing a question to the judge.

Transcript 8 (Judge 21) – In this case on the LHS it was a hidden human entity whereas on the RHS it was the machine JFred. The judge decided that on the LHS it was a machine and on top of that gave the entity a mark of only 20/100 for their human-like

conversation. In other words they thought the conversation was not very much at all like that of a human. Meanwhile the judge was clear that the RHS (the machine JFred) was a male human who was most likely from the USA.

The judge's decision in this case was probably not surprising partly because the conversation with the hidden human was quite terse. The hidden human entity was asked on more than one occasion what their name was to which they replied "I don't know".

Transcript 9 (Judge 29) – The LHS in this conversation was a human whereas the RHS was the machine JFred. The judge decided that the LHS human was definitely a machine and that the RHS machine JFred was an adult non-native English speaking male human. This conversation appears to fall into the same category as Transcript 2 in that the amount of conversation with the hidden human was far less than that with the machine. In fact the judge made only 4 utterances to the human in the 5 minutes available, whereas they made 7 utterances to the machine in the same time period.

That said, in the conversation with JFred there were several responses by the machine which didn't seem to follow from the previous utterance of the judge. This didn't appear to deter the judge though. On the other hand the conversation with the hidden human seemed to be quite dull in comparison with that of JFred and perhaps this was the main reason for the imbalance in utterances exhibited by the judge.

Transcript 10 (Judge 30) – In the last conversation presented in this paper the LHS was the machine Eugene whereas the RHS was a male human. The decision of the judge was that the LHS entity was a male, non-native English speaking teenager whilst the RHS entity was definitely a machine. Indeed they awarded the RHS entity 60 out of 100 for their ability to communicate in a human-like way. So here we see both the Eliza effect in the case of Eugene and the confederate effect in terms of the hidden human.

The human conversation was quite short and seems very disjointed. It is a shame that the judge did not respond more quickly so that the conversation could have had more content. Based on the transcript though, it is not at all surprising that the human was classified as a machine as it appears that they weren't really following the conversation. Eugene's responses were much more to the point and it looks as though the judge enjoyed more conversing with Eugene than with the human. As well as asking the judge questions, the occasional spelling mistake seems to add human credibility.

## MACHINES AND JUDGES

The five machines involved in these tests were: Eugene Goostman, Elbot, JFred, Ultra Hal, and Cleverbot. The dialogue style is dependent on each machine developer's approach to artificial conversation. Four of the five machines, namely Eugene, Elbot, JFred and Ultra Hal are set up to create their 'own' conversations. In the case of Cleverbot however, judges were actually reacting to a 'captured' human response. Cleverbot has been developed to 'can' the responses from humans interacting with the system across the Internet, outputting them elsewhere with context programming. Thus this machine answered a judge's question with a human response against a different input in another conversation somewhere in time. In this experiment the Cleverbot paradigm was the least successful, that is, none of the judges attributed the Cleverbot system with a human classification, this action indicates that the type of context algorithm developed for this machine in itself is not enough to deal with the spontaneity of human conversation. This is an important point to make for future developers.

A wide variety of judges and hidden humans were involved in the tests, both young and old, male and female, etc., in that we were attempting to aim towards Turing's statement concerning "average interrogators" (Turing, 1950). The ploys and strategies of the machines were applicable to all judges however and machines did not exhibit different strategies for different judges, other than those which came out in the normal course of conversation as can be seen in the transcripts included here.

What we have looked at in this paper specifically are cases in which at least one of the pair of hidden entities (either the machine or the human or both) was not correctly identified. Whilst the Turing test itself focusses on the misidentification of the machines involved, the case of humans being misidentified as machines is also of interest. Such cases have been looked at (in terms of previous tests) in depth elsewhere (e.g. Warwick et. al., 2013; Warwick and Shah, 2015). Humans are all quite different and some can exhibit features such as spontaneity, draw interesting relationships, provoke, try to control the discourse and use language, examples and knowledge that a judge may not understand. All these features are likely to assist in a human being misidentified as a machine.

## DISCUSSION

Hayes and Ford (1995) argued strongly that the Turing test leaves a lot to be desired, for example "The imitation game itself has some basic design flaws". Nevertheless they accepted that "exactly what counts as 'cognitive' will shift and change ... be altered by the science itself, just as meanings of words like 'energy' has been changed by physics" (1995: p. 977). Without directly grumbling about the judges, the essence of Hayes and Ford's criticism is that judgement in an imitation game relies on the

subjectivity of the humans as interrogators: “The imitation game conditions say nothing about the judge, but the success of the game depends crucially on how clever, knowledgeable, and insightful the judge is. A clever judge will be looking out for subtle signs” (Hayes and Ford, 1995). So the suggestion here appears to be that the judges should know what they are doing. However, as a direct contrast to this, as we have pointed out, Turing stated clearly that the judges should not be experts

To put this into perspective, it is certainly true that Turing did not pursue the personality of his judges, although he did openly discuss the judges being part of a jury and this indicates a cross section of people from a range of walks of life with some individuals, who are unable to cope or who are otherwise barred, being excluded. Further, Hayes and Ford’s criticism lies in Turing’s imitation game being tied to language: “we should take it as our goal to build something which is just like us?” and “why would anyone want to build machines that could pass the Turing test?” (Hayes and Ford, 1995: p. 974). It is worth remembering that Turing felt that language was an impressive human achievement. The authors observe that communication is the essence of today’s technological and increasingly digital society. Engaging with our machines in natural language is a necessary goal and should not be put off because it is difficult to do.

Turing did specifically state that judges should not be “experts about machines” (Turing et al. 1952). Hence in the case of the Turing test it is questionable just how ‘clever’ judges should be. In the 2012 Bletchley Park tests the two most successful judges, those who correctly determined human from machine, were non-experts: one an adult, the other a young teen, both male. Meanwhile in the tests conducted at the Royal Society in 2014, i.e. those discussed in this paper, all judges had achieved at least a reasonable level of education although the majority could be said to be very well educated, including an FRS and 3 University Professors amongst their number – not really the “average interrogators” that Turing imagined.

It has also been said that “If an artificial intelligence could reliably pass a given instantiation of the test, it would have demonstrated either that its intelligence was genuine or that the judge was not clever enough to ask sufficiently telling questions” (Hayes and Ford, 1995). We totally disagree with this statement on both counts, there is no basis for it whatsoever. The Turing test has never been a test of intelligence (McDermott, 2014), whether genuine or otherwise. Indeed Turing himself spoke against such an idea. Rather than directly tackle the question “Can Machines Think?” Turing initiated his imitation game, the Turing test, as an alternative to explore if a machine could provide satisfactory answers to any questions. As has been witnessed in the transcripts shown here, judges who would quite readily be classified by others as being ‘clever’ made the wrong decision based entirely on a conversation that they instigated.

It certainly has to be acknowledged that the severity of the test depends heavily on the judges taking part. Clearly they are an important feature of the test itself. With a jury of 12 computer science or expert judges it is unlikely that a machine would fool at least 30% of them at the present time. On the other hand, perceptive and non-experts but good conversationalists might also correctly distinguish between human and machine and not fall for the machine banter, as in the case of the best judges in 2012.

However to consider the employment of a human judge to be a design flaw in the test is incorrect. Turing was, we believe, attempting to learn how humans think in order to engineer smart machines. Indeed it is a strong point of the test that it can be instantiated in a very practical way, as can be seen from the discourses tabled, taking into account the representative opinion of several members of the public. It is worth pointing out that the transcripts, and hence the incorrect decisions, presented here arise from one third, 10 of the 30 judges, who took part in the 2014 experiment. In the majority of transcripts (out of 150) there was no error at all in the decisions made, in that the judge in each case correctly identified both the LHS and RHS entity as being either human or machine.

## CONCLUSIONS

More than five minutes is, we feel, overly long a duration to test the current state of technology of artificial dialogue systems. However five minutes is sufficient time to examine how human interrogators come to their decision on what constitutes humanlike or machinelike conversation in an imitation game examining whether a machine provides satisfactory answers to any questions. A valuable feature of Turing's imitation game is not whether a machine gives a correct or incorrect response or a truthful or untruthful one, but rather if it gives the sort of response that a human would give, such that a judge cannot tell the difference (Warwick, 2011; Shah, 2010) between the machine and a parallel human foil. One ploy which can be seen in a number of the transcripts here by machines was that of not directly answering a question but rather attempting to steer the conversation either by a question retort or by changing the subject (Warwick, 2012). This technique often works well in everyday human life and clearly has had a dramatically positive effect here. It is just the sort of thing that humans do.

In this paper we have selected ten specific transcripts arising from ten of the thirty judges in total. It should not be considered that the remaining twenty judges were perfect. On several occasions a judge decided that a hidden human was a machine, at other times a judge was simply unsure about a discourse and at times a different machine was considered to be human. In some cases it is obvious to see why such a selection was made however in other cases we have had considerable difficulty in understanding why a judge made the decision they did.

What we have presented here are cases when a machine was either considered to be human or where the judge was not sure and in each case we have attempted to unravel some of the thought processes involved. Some tactics have been shown by machines to help their success: 1. The machine has a character, 2. It poses questions to the interrogator, 3. It occasionally throws in spelling errors, 4. It occasionally uses humour. However, there is a long way to go before machines can be made to achieve human-like conversation for long periods of time (say 30 minutes or more) and be able to enforce a considerable portion of a jury of human interrogators into believing that they were interacting with another human.

The judge in each case was the main driver of each discourse. It was the judge who started off the discourse and made the final decision. The only role of the hidden entities was in terms of their interaction with the judge. The important decisions made were those by the judge who instigated each discourse. As a reader you may well look over a particular transcript and be amazed that, given the conversation, the judge could make the decision they did. In a sense this is neither here nor there. The machine was not trying to fool you (an individual), it was trying to fool the judges on the day of the test. Similarly for an on line bot to be tested by an individual (perhaps a newspaper reporter) and deemed to be of poor conversational quality is similarly neither here nor there, it is how the machine (which may well have very little of a relationship to the on line bot) performs on the day of the Turing test that matters.

#### REFERENCES

- Bringsjord, S., Bello, P. and Ferrucci, D., "Creativity, the Turing Test and the (Better) Lovelace Test". *Minds and Machines*. Vol. 11, Issue 1: pp. 3-27, 2001.
- Chomsky, N., "Turing on the "Imitation Game"", Chapter 7 in R. Epstein et. al. (eds) *Parsing the Turing test*, Springer, 2008
- Dennett, D. C., "Turing's gradualist vision: making minds from proto-minds", Invited talk: *Turing in Context II*, Brussels, 10 October 2012
- Epstein, R. "The Quest for the Thinking Computer", in R.Epstein, G. Roberts and G. Beber (Eds) *Parsing the Turing test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*", Springer: US pp. 3-12, 2009
- Hayes, P. and Ford, K., "Turing Test Considered Harmful", Proc. Int. Joint Conference on Artificial Intelligence, Montreal, Vol.1, pp.972-977, 1995.
- Khooshabeh, P., Dehghani, M., Nazarian, A. and Gratch, J., "The Cultural Influence Model: When Accented Natural Language Spoken by Virtual Characters Matters", *AI & Society*, DOI:10.1007/s00146-014-0568-1, 2014
- Loebner Prize., "Home of the Loebner Prize" available here: <http://www.loebner.net/Prize/loebner-prize.html>, 1991

- Loebner, H. "In Response to Stuart Shieber's Lessons from a Restricted Turing test" available here: <http://www.loebner.net/Prizef/In-response.html> 1995
- McDermott, D., "On the claim that a table-lookup program could pass the Turing test", *Minds and Machines*, Volume 24, Issue 2, pp 143-188, 2014
- Reidl, M. "The Lovelace 2.0 test for Artificial Creativity and Intelligence". In 'Beyond the Turing Test' 2014 workshop in *Association for the Advancement of Artificial Intelligence 2014*, available here: <http://arxiv.org/pdf/1410.6142v1.pdf>
- Shah, H., "Deception detection and machine intelligence in practical Turing tests" PhD thesis, The University of Reading, 2010
- Shah, H., & Henry, O., "Confederate Effect in Human-Machine Textual Interaction", *Proceedings of 5th WSEAS Int. Conf. on Information Science, Communications and Applications (WSEAS ISCA)*, Cancun, Mexico, ISBN: 960-8457-22-X, pp. 109-114, May 11– 14, 2005.
- Shah, H., & Warwick, K., "Hidden interlocutor misidentification in practical Turing tests", *Minds and Machines*, 20, pp. 441-454, 2010a.
- Shah, H. & Warwick, K., "Testing Turing's five-minutes, parallel-paired imitation game" *Kybernetes*, Vol 39(3), pp. 449-465, 2010b
- Shah, H., Warwick, K., Bland, I., Chapman, C.D., and Allen, M.J., "Turing's Imitation Game: Role of Error-making in Intelligent Thought", *Turing in Context II*, Brussels, 10-12 October, pp. 31-32, 2012. <http://www.computing-conference.ugent.be/file/14> - presentation available here: [http://www.academia.edu/1916866/Turings\\_Imitation\\_Game\\_Role\\_of\\_Error-making\\_in\\_Intelligent\\_Thought](http://www.academia.edu/1916866/Turings_Imitation_Game_Role_of_Error-making_in_Intelligent_Thought),
- Shieber, S., "Lessons from an Restricted Turing test", *Communications of the Association for Computing Machinery*, Vol. 37(6), pp. 70-78, 1994
- The Imitation Game. Weinstein and Black Bear Productions: <http://theimitationgamemovie.com/> 2014
- Traiger, S., "Making the Right Identification in the Turing Test", *Minds and Machines*. 10, pp.561-572, 2000.
- Turing, A.M., Braithwaite, R., Jefferson, G. and Newman, M., "Can Automatic Calculating Machines Be Said to Think?" Transcript of 1952 BBC radio broadcast, in (Eds) S.B.Cooper & J. van Leeuwen, *Alan Turing: His Work and Impact*. Elsevier: Oxford, UK, pp. 667-676, 2013
- Turing, A. M., "Computing Machinery and Intelligence", *Mind*, Vol.LIX (236), pp.433–460, 1950.
- Warwick, K., "Artificial intelligence: The basics", London: Routledge, 2011.
- Warwick, K., "Not Another Look at the Turing Test!", In Proc. SOFSEM 2012: Theory and Practice of Computer Science, Lecture Notes in Computer Science, Vol.7147, M.Bielikova, G.Friedrich, G.Gottlob, S.Katzenbeisser and G.Turan (eds.), Springer-Verlag, pp. 130-140, 2012.

Warwick, K. and Shah, H., "Good Machine Performance in Practical Turing Tests", *IEEE Transactions on Computational Intelligence and AI in Games*, Vol.6, Issue.3, pp.289-299, 2014a

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

Warwick, K. and Shah, H., "Human Misidentification in Turing Tests", *Journal of Experimental and Theoretical Artificial Intelligence*, Vol.27, Issue.2, pp.123-135, 2015

Warwick, K., Shah, H. and Moor, J.H., "Some Implications of a Sample of Practical Turing Tests", *Minds and Machines*, 23, pp.163-177, 2013.