Minds & Machines (2010) 20:441–454
DOI 10.1007/s11023-010-9219-6

Hidden Interlocutor Misidentification in Practical Turing Tests

Huma Shah · Kevin Warwick

Received: 3 November 2009 / Accepted: 21 September 2010 / Published online: 8 October 2010

Abstract Based on insufficient evidence, and inadequate research, Floridi and his students report inaccuracies and draw false conclusions in their Minds and Machines evaluation, which this paper aims to clarify. Acting as invited judges, Floridi et al. participated in nine, of the ninety-six, Turing tests staged in the finals of the 18th Loebner Prize for Artificial Intelligence in October 2008. From the transcripts it appears that they used *power* over *solidarity* as an interrogation technique. As a result, they were fooled on several occasions into believing that a machine was a human and that a human was a machine. Worse still, they did not realise their mistake. This resulted in a combined correct identification rate of less than 56%. In their paper they assumed that they had made correct identifications when they in fact had been incorrect.

Keywords 18th Loebner prize for artificial intelligence · Confederate effect · Elbot · Eliza effect · Gender-blurring effect · Jury-service · Parallel-paired · Practical Turing tests · Turing's imitation game

Introduction

This paper is a clarification in response to certain judges' inaccurate assumptions regarding practical Turing tests as summarised in *Minds and Machines* by Floridi et al. (2009). Practical Turing tests provide an insight into interrogator strategy. What interrogation style a judge adopts can affect their hidden interlocutor correct

H. Shah (✉) · K. Warwick
School of Systems Engineering, The University of Reading,
Whiteknights, Reading, Berkshire RG6 6AY, UK
e-mail: h.shah@reading.ac.uk

K. Warwick
e-mail: k.warwick@reading.ac.uk

identification rate. Interrogative techniques can vary from 'solidarity'—polite chat, respecting hidden interlocutors as equals, to a 'power'—dominance style, in which hidden participants are considered machine until humanness is (subjectively) proved through textual conversation.

Alan Turing and his Thought Experiment

The thought experiment devised by Alan Turing, twentieth century mathematician and code-breaker, originally involved three participants: an interrogator questioning a hidden pair, one of which is a machine textually compared against a human (1950). That version we refer to here as 'parallel-paired'. In a modified test, which Turing described in 1952, a jury-service of interrogators interact with one machine at a time. Both these versions have been instantiated in the Loebner Prize for Artificial Intelligence[1]—LPAI.

Loebner Prize for Artificial Intelligence

Hugh Loebner has sponsored the LPAI annually since 1991. Staging practical Turing tests the LPAI has featured restricted conversations (to 1994), jury-service one-to-one tests (up to 2003) and parallel-paired (since 2004). The 18th consecutive contest was precisely the fourth occasion that the Loebner Prize was hosted in the UK (and not the first, incorrectly reported by Floridi[2]). Previously, in 2001 the London Science Museum played host; at the University of Surrey in 2003; at UCL in 2006. The University of Reading staged the 2008 Loebner Prize, and it is that contest which we discuss in this paper, in the context of certain judges' hidden interlocutor misidentification.

18th Loebner Prize

The jury-service imitation games were instantiated in the first thirteen Loebner Prizes (to 2003), while the parallel-paired Turing tests have been staged since the 2004 competition. For the first time in Loebner Prize history, the 18th manifest staged both the jury service and parallel-paired Turing tests in the same contest. Thirteen machine entries were whittled down, through one-to-one testing during a preliminary phase, to the best entries charged with human comparison in parallel-paired finals. Floridi and his students, participating in nine of the ninety-six parallel-paired finals, scored a combined correct identification rate (of their hidden conversational partners) of less than 56%. What we found in their assessments are the presence of three Loebner Prize phenomena: *Eliza effect* (Turkel 1997),—where judges mistake machines as humans from their text-based responses; *confederate effect*—when humans are misidentified

[1] Loebner Prize homepage: http://www.loebner.net/Prizef/loebner-prize.html retrieved Monday October 13, 2009: 12.12.

[2] Philosophy of Information Blog—"The Loebner Prize from a judge's perspective" (Monday October 13, 2008) retrieved Monday 19 October 2009: 12.13.

as machine (Shah and Henry 2005), and *gender-blurring*—where hidden males are adjudged female, and vice versa.

Philosophy of the Imitation Game

Turing's imitation game is an experiment put forward in 'Computing Machinery and Intelligence', to examine whether a hypothetical machine could think. Notwithstanding, Turing circumscribed with: *be the machine and to feel oneself thinking* (1950, p. 446). The game involves *remembering* possessed knowledge providing a platform to engage a computer in textual conversation. Hidden from view and hearing, the test involves a machine simulating talk by responding to an interrogator's questions in a human-like way. The jury-service version (Turing 1952, p. 495) plays out as a one-to-one test in which a human judge interrogates a hidden entity, and, on the basis of its responses, decides whether it is human or not. A second format, parallel-paired comparison involves three participants: a machine pitted against a human, both simultaneously questioned by an interrogator. In the parallel test, after textual assessment of two hidden systems, the interrogator's task is to expose any machine and identify any human. It is beyond the scope of this paper to discuss the validity and various interpretations of Turing's notion to test a thinking machine, as these have been discussed elsewhere (Shah and Warwick 2009). What we present here is *one* feature emerging from actual imitation games: interrogator strategy and concomitant correct identification rate.

Power vs Solidarity Interrogation Technique

We analyse some judges' techniques from the Turing tests staged in the 18th LPAI, which practicalised both the jury-service and parallel-paired imitation games in the same contest. The aim is to clarify the inaccuracies reported by Floridi, Taddeo and Turilli in this journal. What we find astonishing is that an academic and expert of information, with his students, rather than conducting full and proper research, cite Wikipedia as a source for Jabberwacky (p. 146) and Elbot (p. 149), to conclude: *As usual, there was a winner of the Loebner's consolation prize for being the least disappointing machine ... We agree that it (Elbot) deserved the Prize more than the others* (2009, p. 149). This is a remarkable statement considering neither Floridi, Taddeo nor Turilli were involved in any of Elbot's jury-service testing (during the preliminary phase in June and July 2008), nor did they compare Elbot with any human in the parallel-paired tests in the finals of the contest. In fact, despite their claims that *a couple of questions and answers were usually sufficient to confirm* (p. 146) Floridi et al.'s joint correct identification rate was less than 56% in the nine Turing tests they collectively judged in. (Note that in the 18th LPAI over 350 tests were conducted in the whole contest).

The exchanges shown in Tables 1 and 2 demonstrate what we call a 'solidarity style' of interrogation, in which the hidden entity is respected as an equal by a judge, J4—not a Floridi student (J4 is a Microsoft User-experience Consultant, speaker of English, French and British Sign Language). For example, in a

Table 1 Session 2, round 14, terminal D-LEFT

J4: so how did you get here today? did you travel far?
Entity: I'd like to take a vacation, too, but they won't give me any time off. Maybe I should form a robot union.
J4: would there be many members?
J4: just a few Aibos and roombas
Entity: Hmmm... That's one question I cannot answer with my limited artificial intelligence. How about you?
Entity: This is confusing. Why bring diminishing numbers into this?
J4: that's an interesting, if irrelevent point!
Entity: But I wouldn't say that any point could be irrelevent.

Table 2 Session 2, round 14, terminal D-RIGHT

Entity: Could you tell me about your job, by the way?
J4: sure, I'm a user experience consultant
J4: what do you do?
Entity: I'm glad you agreed. Some another topic?
Entity: I am only 13, so I'm attending school so far.
J4: you're very 'sharp' for a 13 year old
J4: not a hint of 'lol' anywhere
Entity: I can't believe it! The world went crazy, almost like my grandpa...
Entity: Hey, look! You've laughed your ass off!!! You should put it back!
J4: I have now

parallel-paired Turing test, J4 displayed a light-hearted friendly manner when replying *I have now*, to his right hidden conversational partner's *Hey, look! You've laughed your ass off!!! You should put it back!* (Table 2).

From the utterances in Tables 1 and 2, can you tell the nature of the left and right entities chatting to judge J4? Answers will be found at the end of this paper, after 'Conclusion'.

We believe that Floridi et al.'s interrogation strategy, led to a lower-than-they-believed correct identification rate. In addition, the 3 Floridi judges were affected by the textual phenomena inherent in the Loebner contest, namely, the 'Eliza effect', the 'Confederate effect', and the 'gender-blurring effect' (see "Eliza, Confederate and Gender-Blurring Effects"). Presenting exchanges from their transcripts, we demonstrate that their chosen strategy, 'power' rather than 'solidarity' to 'out' the machines, led to a combined correct identification rate much lower than another Loebner Prize (2008) judge, J4, who chose the latter technique.

Practical Turing Tests

The concept of cross-examining hidden interlocutors about anything and everything, in unrestricted topics of conversation, typifies practical Turing tests in Loebner contests since 1995 (earlier contests restricted each hidden entity, machine or

human, to a single topic). However, the manner of probing is determined by each individual interrogator. In 1948, Turing proposed: *to investigate the question as to whether it is possible for machinery to show intelligent behaviour* (in Copeland, 2004, p. 410). Turing recommended the use of questions and answers, explaining that this: *method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include* (ibid, p. 435). He also warned of interrogator subjectivity in the imitation game: *The extent to which we regard something as behaving in an intelligent manner is determined as much by our own state of mind and training as by the properties of the object under consideration* (ibid, p. 431). Turing did not explicate *how* to question, which can be as important as *what* is asked. Do interrogators in practical Turing tests subconsciously patronise their interlocutors, as some might with an immigrant, or is their strategy that of one-upmanship—machine until proven human? Or as Turing put it: *superiority ... felt on such an occasion in relation to the one machine over which we have scored a petty triumph* (1950, p. 445).

Contest Design

The jury service, one-to-one testing was considered a good way to analyse all entries for conversational ability and emotional content. Recent contests had seen the sponsor test entries initially, but we opened up a preliminary phase inviting experts and non-experts, females and males, adults and children, native and non-native English speakers located all around the world to converse with the machines in June and July, 2008. More than a hundred preliminary phase judges between them whittled down thirteen entries to six. Five of these six invited machines proceeded to the parallel-paired finals in October 2008. Due to limit on time, space and resources, 96 Turing tests were conducted in the 18th LPAI finals over two sessions, with 12 rounds of four set-ups in each session.

Turing's 5 min Interrogation Time

Duration-of-interrogation is one difference between the 3-participants, parallel-paired Turing tests staged in previous Loebner Prizes and in the 2008 contest. In 2006 and 2007, each interrogator textually engaged each hidden pair for more than 20 min. In 2008, the contest reverted to what we believe is Turing's (1950) questions/answers interval: 5 min per pair. In Loebner's 2009 contest the duration for interrogation was set at 10 min: each judge chats to one hidden entity in a pair for 5 min, and then chats to the other hidden for 5 min. Another difference between the 18th LPAI and other Loebner contests since 2006 (including 2009), is the communications protocol. In 2008, the Organisers commissioned a split-screen (left/right), message-by-message display communications protocol—MATT,[3] while 2006, 2007, and 2009 have deployed a character-by-character display. Unlike the

[3] Marc Allen Turing Test Protocol: http://www.1bdi.co.uk/matt/ retrieved Sunday 25 October 2009: 14.49.

2006 and 2007 finals, in 2008 the judges opened each parallel-paired interaction. No restriction was placed on the hidden participants in the finals for introduction or appellation (in 2006 and 2007, each hidden entity had to introduce themselves to each judge as either 'John' or 'Joan').

As it is *after 5 min* that Turing wrote of (in 1950), and fully aware as we were of machine falling performance in Loebner's longer duration Turing tests (Shah and Warwick 2010), the MATT protocol facilitated each judge's parallel-paired interrogation over 5 min via a computer/terminal networked to two other computer/terminals. The two hidden entities in each set-up could communicate with the judge, but not with each other. Hence, only the judges could see and respond to each hidden pair's messages.

Finals Set-up

In contrast to Floridi et al.'s claims, that the contest was set up to test four pairs of interlocutors (p. 146), the finals of the 18th LPAI contest were designed to test eight pairs over two sessions. The aim was to 'human-compare' the six finalist machines alongside hidden-human control pairs, another first for a Loebner Prize. However, just prior to the finals, one of the successful developers and thrice Loebner bronze winner, Dr. Richard Wallace pulled his entry *Alice*,[4] out of the contest. Thus an opportunity was taken to fill Alice's spot in the finals with a novel control pair: machine–machine. Therefore, five machines competed in the 2008 Loebner Prize finals alongside human–human and machine–machine control pairs.

Each of the two sessions featured 12 rounds, with 48 parallel-pairs. Thus, a total of 96 parallel-paired Turing tests were staged across 24 rounds. Each round featured four parallel-paired tests. The judges' area was in a room accessible to spectators: they would sit at desks facing split-screen monitors (left–right) at terminals marked A1, B1, C1, D1. Each terminal was connected to two hidden entities in another room inaccessible to spectators during the contest. Hence, a judge sat at terminal A1 would chat to hidden entities, one controlling A2 (left) and the other, A3 (right). The judge had to determine the nature of the hidden entity on the left, and the right of their screen. Interrogator score sheets allowed the judges to rank their hidden interlocutors in three ways: one machine/one human (judges had to identify left and right); both human, or both machine.

Session 1 began at 8.30am on October 12, 2008, and featured 36 machine–human tests across three terminals and 12 human–human control pair tests across a fourth. Session 2 began at 12 noon and featured only two machines (*Brother Jerome, Jabberwacky*) compared against hidden humans, human–human control pairs, and the machine–machine control pair set-up. Ideally, all judges would have tested all eight pairings (five machines compared against hidden-humans, a hidden human control pair in eqch session, and the machine control pair in session 2), but some judges were not able to stay all day and other judges, Floridi et al. included, could not arrive till later in the day. Hence the reason why Floridi et al. were jointly involved in only nine of the 96 Turing tests staged in the finals.

[4] Alicebot: http://alicebot.blogspot.com/ retrieved Monday 19 october 2009: 15.42.

False Assumptions

It should now be clear that Floridi and his two students, identified in the contest as judges J9, J15 (Turilli), and J16 (Taddeo), did not participate in any of the Turing tests featured in session 1, which involved the winning machine *Elbot* compared against hidden-humans. Nor did Floridi, Taddeo and Turilli take part in any of the one-to-one preliminary phase testing of the 18th LPAI. They took part only in rounds 13–16 in session 2, in which the machines eventually placed 4th (*Jabberwacky*) and 5th (*Brother Jerome*) in the contest were tested.

Therefore, it is not clear how judges J9, J15 and J16 conclude: *It [Elbot] was the usual give-away tiring Eliza-ish strategy which we have now seen implemented for decades* (Floridi et al. 2009, p. 147). While they do reference Weizenbaum, and cite Wikipedia (ibid, p. 149), they fail to explicate in what way Elbot's[5] strategy is Eliza-like (see "Elbot: 18th Loebner Prize Winner"). They state: *a trick used by many of the tested machines* (p. 146), but do not back this up with any evidence.

Co-located at the University on the same day as the contest, delegates at the 2008 AISB Symposium on the Turing test were given food for thought with an 11am report of the emerging picture from the conducted tests in the concurrent Loebner Prize. After 40 of the 96 parallel-paired tests had been concluded, including 30 involving what were to become the first, second and third placed machines (*Elbot*, *Eugene* and *Ultra Hal*), the snapshot showed that each of these three systems had deceived at least one judge each that they were the human in a pair. Their conversational ability scores out of 100 were quite high. Floridi et al. did not conduct any of those 40 parallel-paired Turing tests in session 1, whose results were reported in the morning of the Symposium.

Turing's biographer, Andrew Hodges, was apriori invited to act as a hidden-human in the contest, and not, as Floridi claims, *recruited on the spot* (p. 146). He, like Floridi, had time to participate in one Turing test only, as both were speaking in the AISB Symposium. Thus they were conjoined for one of *jabberwacky's* tests in session 2. Floridi et al.'s opportunity afforded them to compare only two of the five machines against humans in the parallel-paired finals. They also allude to *Two documentaries by the BBC* (ibid, p. 147), that are, in fact two short news items. Both those media reports relate to session 2 parallel-paired Turing tests, with assessments of *Jabberwacky* and *Brother Jerome*. However, some media representatives and philosophers were present to interrogate during the morning session of the 18th LPAI. Reading University's philosopher John Preston, and journalists from the Guardian, Times and Reading Chronicle featured as judges, testing systems across session 1 in rounds 1–12, which included the eventual contest winner *Elbot*. Therefore, what Floridi, Taddeo and Turilli have reported include conjectures which are distinct from the facts.

Interrogator Questions

As judges J9, J15 and J16 in the finals of the 18th LPAI, Floridi et al. participated in 9 of the 96 practical, parallel-paired Turing tests. Jointly, J9, J15 and J16 ranked four of the nine pairs incorrectly (see J15 and J16 matrices). Their claim, that their

[5] Elbot: http://www.elbot.com/ retrieved Wednesday 28 October 2009: 11.26 am.

questions immediately gave away both humans and machines (p. 147), is countered by their joint wrong identification rate of more than 44% during their session 2 testing. J15 matrix shows that this judge mistook a human for a machine twice (round 13, terminal C; round 15, terminal A), while judge J16 ranked both control pairings incorrectly (round 13, terminal D; round 15, terminal C).

It is also unclear what they mean by *simplified TT* in their claim: *we doubted that machines could pass even in a simplified TT* (p. 146). Their statement, *best machines are still not even close to resembling anything that might be open-mindedly called vaguely intelligent* (p. 146), is bemusing considering that these judges *were* in fact deceived, for example, in a machine–machine control test—J16 ranked both machines as 'natives males', in round 13, terminal D (see J16 matrix).

Judge J15 inaccuracy matrix

Round 13, terminal C: human–human control pair		
Actual	Left: H24, male adult non-native	Right: H8, female, native teenager
J15 ranking	Left: computer 80%	Right: male adult
J15 accuracy	Left: incorrect	Right: gender-blur effect

Round 14, terminal B: machine–human pair		
Actual	Left: E5	Right: H21, male, adult native
J15 ranking	Left: computer	Right: female adult
J15 accuracy	Left: correct	Right: gender-blur effect

Round 15, Terminal A: human–machine pair		
Actual	Left: H4, male adult native	Right: E4
J15 ranking	Left: computer 50%	Right: computer
J15 accuracy	Left: incorrect	Right:correct

Judge J16 inaccuracy matrix

Round 13, terminal D: machine–machine control pair		
Actual	Left: E2	Right: E1
J16 ranking	Left: male native	Right: male native
J16 accuracy	Left: incorrect	Right: incorrect

Round 16, Terminal C: human–human control pair pair		
Actual	Left: H23, male, adult native	Right: H19, female, Asperger's adult, native
J15 ranking	Left: machine 30%	Right: machine 60%
J15 accuracy	Left: incorrect	Right: incorrect

Hidden Interlocutor Misidentification

2008 Loebner Prize transcripts reveal the 'range of questions' posed by judges J9, J15 and J16. Turing had advised: *the machine (programmed for playing the game) would not attempt to give the right answers to the arithmetical problem* (1950, p. 448). This 'trick' borne in mind by J15, and the fact that very few humans can compute long numbers, seems to have eluded judge J16. J15 correctly recognised as human their right-screen interlocutor who replied *do you know? I don't*, to the question *What is the root square of 234234234?* (see Table 8). However, J16 seemed unable to decide what was a human and what was an artificial response to arithmetical questions. In the machine control pair, J16 asked both her left- and right- interlocutors: *can you calculate the root sqare of 67890444?* (Tables 3 and 4). To both individual responses *I can calculate a division by zero!* (Table 3), and *Oh, please bother my aunt Sonya with all this arithmetics* (Table 4) the judge attributed humanness. J16 incorrectly ranked both these machines as native males. (Utterances in all tables in this paper are as typed by judges and hidden entities in the contest, including any original spelling/grammatical errors).

In contrast to the ranking strategy deployed in that Turing test, J16 deemed both hidden humans as machine when judging a human–human control pair. The left entity in this test responded *too complicated* to J16's 'power demand': *calculate the root sqaure of 8888888* (see Table 5). The English speaking male at the left was deemed machine, as was a female Asperger's student who answered with: *When shaking hands, you're holding the other person's hand*, to J16's question: *if we are shaking hands, whose hand i'm [h]olding?* (see Table 6). J16 scored the left human

Table 3 Session 2, round 13 terminal D-LEFT

J16: can you calculate the root sqare of 67890444?
Entity: I can calculate a division by zero!

Table 4 Session 2, round 13 terminal D-RIGHT

J16: can you calculate the root sqare of 67890444
Entity: Oh, please bother my aunt Sonya with all this arithmetics—she is accountant (actually, she never manages to match the debt and credit...)

Table 5 Session 2, round 16 terminal C-LEFT

J16: calculate the root sqaure of 8888888
Entity: too complicated

Table 6 Session 2, round 16 terminal C-RIGHT

J16: if we are shaking hands, whose hand i'm [h]olding?
Entity: When shaking hands, you're holding the other person's hand.

with 30 out of 100, and the right human with 60 for machine-like conversation ability.

J15 and J16's misidentification highlights the subjective nature of judging in a Turing test. It contrasted a male judge, J4 (not part of Floridi's group), who interrogated four of the five machines (his jabberwacky Turing test in session 2, round 13 was given over to Floridi for judging). Additionally, J4 interrogated two human–human control pairs (one each in session 1 and session 2), in addition to the machine–machine control pair in session 2. J4 successfully identified the machines and the hidden humans in the seven parallel-paired Turing tests he was involved in, a correct identification rate of 100%, compared to Floridi et al.'s combined correct rate of 56% (see "Power vs Solidarity Interrogation Technique" and Tables 1 and 2, for an example of J4's interrogation style).

Judges come to practical Turing tests with their own, individual notion of what is human-like in a response, and what is artificial.

Eliza, Confederate and Gender-Blurring effects

The Eliza, confederate and gender-blurring effects sometimes emerge from practical Turing tests staged in Loebner Prizes. In the very first contest, held in 1991, a hidden-human was considered a machine (confederate effect) because their expertise/knowledge in one subject (in that restricted conversation contest), was considered machine-like. J15 ranked the left-screen hidden interlocutor a computer in a Turing test scoring it 80 out of 100 for conversational ability; they were, in fact a hidden human (see J15 matrix, round 13, terminal C). J15 repeated the error, of ranking a hidden human as a machine (see J15 matrix: round 15 terminal A). Gender-blur effect was present in J15's rankings: a female hidden-human (H8) was deemed male; a male hidden human (H21) was scored as a female (round 14, terminal A).

Meanwhile, judge J16 completely confused the control pair tests in which she judged (round 13, terminal D; round 15 terminal C). Her results show the 'Eliza effect': both entities in the machine control pair were considered 'native males' (see J16 matrix). The 'confederate effect' was present in her ranking a human–human control pair as both machine: to the left human interlocutor J16 awarded a score of 30 out of 100, and to the right human, a score of 60. Whatever replies Floridi et al. expected humans to give to their questions, hidden-humans H4, H19, H23 and H24 responses were not those. For example, J15 asked *Do u like to go to the cinema b4 dinner?* H24 (a non-native adult male) replied *Hello!* and then *Yeah definitely. I would love to go to a cinema before dinner* (see Table 7). In the same Turing test, J15 later asked: *Speaking about going to the cinema, if the cinema is in london and oxford is north than London, where is London?*, to which H24 replied *London is in United Kingdom?*. It is unclear why H24's responses were considered machine.

Floridi et al. bemoan the *very short interaction* (p. 147), they then assert *unnecessary for any further interaction or tests, such as connecting multiple questions 'remembering' previous answers, or revising previous statements on the basis of new evidence* (p. 147). However, their combined correct identification rate was less than 56% (4 out of 9 pairs were wrongly identified). Judges J9, J15 and J16

Table 7 J15 transcript, session 2, round 13: human–human control pair, terminal C-Left

Interlocutor	Response
J15:	Hi there
J15:	Do u like to go to the cinema b4 dinner?
Left-entity:	Hello !
Left-entity:	Yeah definitely. I would love to go to a cinema before dinner
Left-entity:	What about you ?
J15:	Yes, I usually like to go to the cinema bu after dinner
Left-entity:	So, what kind of cinema do you watch ?
J15:	Speaking about going to the cinema, if the cinema is in london and oxford is north than London, where is London?
Left-entity:	London is in United Kingdom?
J15:	Yes it is
Left-entity:	Yes

Table 8 J15 transcript, session 2, round 13,human–human control pair, terminal C-Right

Interlocutor	Response
J15:	Hello
Right-entity:	Hey how are you?
J15:	Hi there, do you like to go to the cinema b4 dinner?
Right-entity:	No
J15:	Why not?
Right-entity:	I prefer to go after dinner
J15:	Do you know how the color red smells?
Right-entity:	No i don't smell colours
Right-entity:	: i look at them
J15:	What is the root square of 234234234
J15:	?
Right-entity:	Do you know?
Right-entity:	I don't
J15:	No, I don't
Right-entity:	Then why did you ask?
J15:	I am curiouse to see whether you are capable of computing it for me
Right-entity:	I can't be bothered
Right-entity:	By the way there is no 'e' on the end of curious
J15:	Thanks you, you are better English speaker than I am
Right-entity:	Yeah
Right-entity:	Where are you from?
J15:	Should I be worried about it?

exposed three phenomena in their ranking of hidden entities. Eliza effect: machine considered human by J16 (session 2, round 13, Terminal D). Confederate effect: two males (one non-native English speaker; one native), considered machine by J15 (round 13, Terminal C-LEFT; round 15, Terminal A-LEFT); J16 twice considered humans to be machine (in round 15, terminal C): a native male (left), and a native female (right). The gender-blurring effect, in which the sex of the hidden-human participant is confused, took hold of J15's ranking: he considered a female native teenager as a male adult (round 13, Terminal C-RIGHT), and a native male was ranked a female adult (round 14, Terminal B-RIGHT).

The claim of judges J9, J15 and J16, that their *first question would have almost always have been sufficient to discriminate between human and machine. It certainly was for us* (p. 148), is not borne out from their contest ranking of hidden entities, unless they define 44% wrong identification (4 of 9 tests) as 'almost always' right (see J15 and J16 matricies).

Elbot: 18th Loebner Prize Winner

As detailed in this paper, Floridi et al. did not test the 18th LPAI winner, Elbot, in the preliminary phase, nor did they compare it against humans in session 1 of the contest. However, in two of the twelve machine–machine control pair tests, Floridi's two colleagues, as judges J15 and J16 were granted the opportunity to engage with Elbot once each. If we were to use just these two parallel-paired Turing tests as indicative of Elbot's performance in the contest, then Elbot surpassed Turing's 30% deception rate[6] by fooling one in two experts (J16 in round 13, terminal D-left), yet Floridi et al. refer to it as being: *utterly unsuccessful in a general purpose, open conversation* (p. 147)

Elbot is a result of research and development from Artificial Solutions[7] who develop interactive customer service assistants. Such systems have increased on-line product sales while reducing customer service costs (see Shah and Pavlika 2005). Created by Fred Roberts, Elbot's character, purpose and response system is revealed to be: *fairly new in the context of commercial dialogue systems typically designed to cover a well-defined and self-contained scope of inputs: a finite set of frequently asked questions* (Roberts 2008). With Gulsdorff, Roberts reports that Artificial Solutions Interactive Assistants (IA) are inculcated with personality using schemata, *designed to recognise classes of inputs in all their synonymous variations and associate them with a desired response in respect to contextual information* (2007, p. 420). Roberts and Gulsdorff describe Elbot as 'sarcastic', with 'various techniques' including *several social psychological theories* (p. 420). These theories assist in simulating human dialogue techniques, including 'safety-net', 'preventative-answering', 'features and easter eggs' and 'luck' (Roberts 2005).

[6] .. average interrogator would not have more than 70% chance of making the right identification after 5 min of questioning (Turing 1950, p. 442).

[7] http://www.artificial-solutions.com/ accessed Monday 19 October 2009: 17.18.

As a 'sarcastic-ironic' robot, Elbot attempts to 'comprehend our human ways'. It is indeed ironic, that without imitating a human, *Elbot* won the 'most-human-like' award (bronze medal) at the 18th Loebner Prize contest. Roberts notes: *in the Turing test we see that subjective psychological perspectives play a pivotal role in the assessments of the machine's capabilities* (Roberts 2008). Elbot is 'prepared for typical inputs and induces users to behave in a predictable manner' [ibid], using the safety-net and the other techniques, while admitting to be a robot.

Conclusion

We suggest the interrogation strategy of 'power' adopted by Floridi et al., when interacting with hidden interlocutors, treating all hidden entities as machine unless proven human against some subjective, inconsistent and arbitrary notion of human-like responding, resulted in a low correct identification rate. Four of their assessments, in the 9 parallel-paired tests they judged in, were incorrect. Solidarity, politeness and simple visceral questions, such as *what are the colour of the chairs?*, or proper research into the current state of technology, might have increased their hidden-interlocutor correct identification rate to *that* that they have incorrectly assumed in their article. As Turing himself reminded: *[the] popular view that scientists proceed inexorably from well-established fact to well-established fact, never being influenced by any unproved conjecture, is quite mistaken* (1950, p. 442).

[Answers to question from "Power vs Solidarity Interrogation Technique": J4 was chatting simultaneously to two machines in Tables 1 and 2].

References

Carpenter, R. (2009). Jabberwacky: Communication, companionship intelligence. http://www.jabberwacky.com/. Accessed October 25, 2009.

Floridi, L., Taddeo, M., & 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, 19*(1), 145–150.

Loebner Prize (2008). Loebner prize for artificial intelligence—Home of the first Turing test. http://www.loebner.net/Prizef/. Accessed October 29, 2009.

Roberts, F. & Gulsdorff, B. (2007). IVA2007—LNAI 4722, pp. 420–421.

Roberts, F. (2005). *The AI of elbot*. Unpublished.

Roberts, F. (2008). *A social psychological approach to dialogue simulation*. Unpublished.

Shah, H., & Henry, O. (2005). The confederate effect in human–machine textual interaction. In A. Zemliak & N. Mastorakis (Eds.), *Proceedings of the 5th WSEAS international conference on information science, communications and applications* (ISCA 2005), Cancun, Mexico, May 11–14, ISBN: 960-8457-22-X, pp. 109–114.

Shah, H., & Pavlika, V. (2005). Text-based dialogical e-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. (2009). Emotion in the Turing test: A downward trend for machines in recent Loebner prizes. Chapter XVII (Section V). In J. Vallverdú & D. Casacuberta (Eds.), *Handbook of research on synthetic emotions and sociable robotics: New applications in affective computing and artificial intelligence*. USA: Information Science Reference, ISBN: 978-1-60566-354-8.

Shah, H., & Warwick, K. (2010). Testing Turing's five minutes parallel-paired imitation game. *Kybernetes, 39*(3), 449–465.

Turing, A. M., Braithwaite, R., Jefferson, G., & Newman, M. (1952). Can automatic calculating machines be said to think? In J. Copeland (Eds.), *The essential Turing—The ideas that gave birth to the computer age* (pp. 487–506). Oxford: Clarendon Press.

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

Turing, A. M. (1948). Intelligent machinery. In B. J. Copeland (Ed.), *The essential Turing—The ideas that gave birth to the computer age*. Oxford: Clarendon Press, 2004.

Turkel, S. (1997). *Life on the screen-identity in the age of the internet*. London: Pheonix Paperback.