AISB 2011

Towards a Comprehensive Intelligence Test

Editors:
Dimitar Kazakov &
George Tsoulas

THE UNIVERSITY of York

Foreword from the Convention Chairs

The AISB'11 call for symposium proposals particularly encouraged events drawing more strongly on the cognitive science aspect of the AISB remit. The result is a coherent programme with a very strong interdisciplinary character, which is also matched in the choice of plenary speakers. The three symposia looking at the interaction between Computing and Philosophy, the prospect of machine consciousness and the quest for a new, comprehensive intelligence test, form a coherent unit where the eternal questions of who we are and what makes us so are asked from a dual Human-Machine perspective. The Symposia on Active Vision, Computational Models of Cognitive Development and Human Memory for Artificial Agents demonstrate how better understanding of the nature and basis of cognitive processes can advance work on Artificial Intelligence and, inversely, how computational models of these processes can help better to understand them. The prominent multi-agent design and modelling paradigm links the Symposium on Social Networks and Multi-agent Systems with the one on AI and Games. Finally, the Symposium on Learning Language Models from Multilingual Corpora, which brings together some of the first attempts in this area, can also be seen through the prism of such a general notion in Philosophy and Linguistics as semiosis, and the dual role of sign and interpretant that text plays in translations.

We are delighted that after another ten successful years in its long history, the AISB convention is returning to the University of York. The 2011 convention takes place on the brand-new Heslington East campus, the result of a multi-million pound expansion that is now the new home of the Department of Computer Science, and hosts the Excellence Hub for Yorkshire and Humber, a new incubator for interdisciplinary research and interaction between academia and industry. The last few years have seen a strong involvement of the Computer Science Department in such interdisciplinary collaboration through the York Centre for Complex Systems Analysis (YCCSA), and we hope that this convention will provide a boost for more synergy between York departments, with other institutions conducting AI-related research in the region, and beyond. As the programme shows, we have also made an effort to promote cooperation with industry and use the convention to support school outreach. The convention format makes it perfect for establishing dialogue and collaboration in new areas of research, as well as across disciplines, and we hope that this year, it will play again this role to the full. We want to thank everyone who has contributed to it or otherwise made this event possible and wish all participants a fruitful and enjoyable time in York.

Dimitar Kazakov and George Tsoulas

PREFACE

2010 marked the 60th anniversary of the publication of Turing's paper, in which he outlined his test for machine intelligence. Turing suggested that the possibility of genuine machine thought should be replaced by a simple behaviour-based process in which a human interrogator converses blindly with a machine and another human. Although the precise nature of the test has been debated, the standard interpretation is that if, after five minutes interaction, the interrogator cannot reliably tell which respondent is the human and which the machine then the machine can be qualified as a 'thinking machine'. Through the years, this test has become synonymous as 'the benchmark' for Artificial Intelligence in popular culture.

New advances in cognitive sciences and consciousness studies suggest it may be useful to revisit this test, which has been done through number of symposiums and competitions. However, a consolidated effort has been attempted in 2010 in the first TCIT symposium. This symposium is a continuation on this effort in a three years project to revisit, debate, and reformulate (if possible) the Turing test into a comprehensive intelligence test, or suite of tests, that may more usefully be employed to evaluate 'machine intelligence' at the dawn of the 21st century.

This is a 3 year project. It was conceived in 2009 by Aladdin Ayesh and supported by AISB committee members, notably Mark Bishop and John Barnden. It was announced during the General Annual Meeting of the AISB 2009. Later on that year it gained support and funding from U.S. Office of Naval Research Global (ONRG).

TCIT 2010 was the first symposium denoting the first stage of this project and focused on establishing the objectives and criteria for the development of a new machine intelligence test to move the interpretation of the Turning Test forward into the 21st Century.

TCIT 2011 is the second stage in which competitions and tests are considered while expanding on some of the issues discussed or emerged from TCIT2010.

TCIT 2012 will be hosted as part of the joint AISB/IACAP conference in commemoration of Alan Turing 100′ s birthday.

ORGANIZING COMMITTEE
Aladdin Ayesh (De Montfort,Symposium Chair)
Mark Bishop (Goldsmith College, London)
John Barnden (University of Birmingham)
Luciano Floridi (Hertfordshire/Oxford)
Kevin Warwick (Reading)

PROGRAM COMMITTEE
Selmer Bringsjord (Rensselaer Polytechnic Institute)
Bernd Carsten Stahl (De Montfort)
James Moor (Dartmouth College)
John Preston (Reading)
Ray Tuner (Essex)
Robb Wilcox (ONRG)

ATTENDEES-TCIT2010 ROUNDTABLE
Doug Samuelson, InfoLogix, Inc.
Prof. Steve Torrance, Sussex University
Dr. Marc Schroder, DFKI GmbH
Dr. Darren Abramson, Dalhousie University
Aimee Reichert, University of Washington
Hugh Loebner, Founder of Loebner Prize
Dr. Rafal Rzepka, Hokkaido University
Prof. Mark Bishop, Goldsmith College
Prof. Murray Shanahan, Imperial College
Dr. Rob Wilcox, ONRG
Prof. Drew McDermott, Yale University
Prof. John Barnden, Birmingham University
Dr. Ed Keedwell, Exeter University
Dr. Aladdin Ayesh, De Montfort University

TABLE OF CONTENTS

Turing's misunderstood imitation game and IBM's Watson success

Huma Shah*

Abstract. At the heart of Turing's 1950 imitation game is the question-answer test to assess whether a machine can respond with satisfactory and sustained answers to unrestricted questions. In 1966, Weizenbaum's Eliza system made it possible for a human and a machine to communicate via text in question and answer sessions. Forty five year later in 2011, IBM Watson achieved remarkable success winning an unrestricted question-answer exhibition match competing against humans in *Jeopardy!* a US TV quiz show. Is it now time to scale up to Harnad's Total Turing Test combining natural language with robot audio and vision engineering?

1 INTRODUCTION

Turing's imitation game is misunderstood to the extent that it has been considered harmful and a burden to the field of artificial intelligence, the science inspired by it [1]. The author argues that Turing's idea to examine machine thinking is unfairly judged by the performance of systems in one instantiation, the Loebner Prize for Artificial Intelligence [2, 3]. According to Levesque, the entries in Loebner's interpretation of Turing's question-answer test tell us nothing about intelligence [4]. In the 45th anniversary year since the 1966 unveiling of Weizenbaum's Eliza system [5], the first computer programme that allowed interaction between human and machine using text-based question-answer sessions, IBM showed in 2011 that its Watson technology [6] is capable of sophisticated natural language processing allowing it to compete against, and beat expert level humans in a general knowledge question-answer US TV quiz show, Jeopardy! [7]. Is it now time to scale up to Harnad's robot Total Turing Test [8]?

This paper aims to extend an understanding of Turing's ideas on machine thinking by considering points from his work between 1947 and 1952. The author contends that there is one imitation game of five minutes duration, allowing a *first impression* [9] and *thin slice of behaviour* [10] to suffice for the assessment of *satisfactory and sustained* answers from a machine to unrestricted questions. Further, Turing's game can be practicalised in two different ways: a) 2-participant, machine directly questioned by a jury member and, b) a 3-participant simultaneous comparison of a hidden machine with a hidden human, both questioned by interrogators. Finally, the author believes IBM Watson has shown that it is possible for a machine 'to think', i.e., to receive and analyse an input and respond with an appropriate output at the same, if not faster speed than a human.

* School of Systems Engineering, The Univ. of Reading, RG6 6AH, UK
Email: h.shah@reading.ac.uk

Watson demonstrated its thought processes under the world's glare during its successful performance in a question-answer *Jeopardy!* exhibition match held over three days in February 2011. The next step involves combining Watson technology with robot engineering to build machines augmenting human life.

In the following section, the author presents Turing's own ideas to examine whether a machine can think. Next, in section 3, Hayes and Ford's critique of the imitation game is reviewed. In section 4, the performance of IBM Watson is analysed. Section 5 summarises and concludes with directions for future research.

2 THE IMITATION GAME

In 1947, Turing gave a lecture to the London Mathematical Society in which he raised the prospect of an intelligent machine competing with a human, pointing to a game of chess as adequate for an initial encounter [11]. In 1948, in a report on 'Intelligent Machinery' [12] Turing used *imitation* for the first time and set the scene to investigate, "whether it is possible for machinery to show intelligent behaviour*"* (p. 410). Turing also discussed *interference* to modify the machine, a process which he considered analogous to a human's modification as a result of learning something new. Turing felt positive about "believing in the possibility of making thinking machinery" (p.420). He added presciently, "further research into intelligence of machinery will probably be very greatly concerned with 'searches' ..." (p. 430). Searching through diverse information of text and numbers helped IBM's Watson machine achieve success in the man vs machine *Jeopardy!* match [6,7].

In his 1948 paper, Turing introduced the forerunner of his 1950 3-participant test in which a machine is compared with a human. In the 1948 version of the imitation game, a mathematician and chess player acted as 'B' operating a 'paper machine', while two 'poor' chess players, A and C, unseen to each other, played across two rooms. Turing felt it might not be easy for C to say whether they were playing A or the paper machine: "C may find it difficult to tell which he is playing" (p. 431). In 1958, Newell and Simon predicted that a computer would become world chess champion within ten years [13]. It may have happened later than predicted, but an IBM machine, Deep Blue did beat a world chess grandmaster in 1997 [6].

In the historic and important 1950 paper, *Computing machinery and intelligence* [14] Turing replaced chess, and the question of whether a machine could show intelligent behaviour, with text-based question-answer sessions to examine whether a machine could think. [In 1948, Turing had written, "the idea of 'intelligence' is itself emotional rather than mathematical" (p. 411)]. Whether the machine could think was to be determined by the machine replying satisfactorily to unrestricted questions

(p.447). Turing claimed that the "question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour" (p. 435). However, Turing did point out the limitations of the machines at that time: "there will be some questions to which it will either give a wrong answer, or fail to give an answer at all however much time is allowed for a reply" (p. 444). Turing wrote "I am often wrong, and the result is a surprise for me" (p. 451), but, he asked, would it be fair to deem machines *worse* for *not* making mistakes? (p. 448). Turing supposed that closed questions, with 'yes' or 'no' answers, were appropriate rather than the type of questions machines would fail to answer coherently, for instance, open question eliciting an opinion or visceral description. Turing felt it might be difficult for a machine to answer a question such as "What do you think of Picasso?" (p.445). Turing reminded "it has only been stated, without any sort of proof, that no such limitations [on answering open questions] apply to the human intellect" (ibid), such as when humans do not have an opinion on a topic, or do not have access to a piece of knowledge.

In the earlier sections of the 1950 paper, Turing introduced and discussed a 3-participant scenario in which a hidden man was compared with a hidden woman. Both hidden humans were to be questioned by a human interrogator who attempted to distinguish the man from the woman, from their text-based answers to questions. By the end of section 5 (p. 442), Turing's machine test was revealed as a version of the human-human imitation game having a digital machine simultaneously compared with a man, the questioner attempting to determine which is the natural and which is the artificial. In section 6, considering "contrary views" to his machine thinking game, Turing dismissed possible objections and evolved his 3-participant comparison test into a direct questioning of a machine through a viva voce style interview (p. 446). It is the 2-participant test, the direct questioning of a machine 'witness' by an interrogator, which Turing elaborated upon during a 1952 radio discussion [16].

In 1979, in an experiment conducted by Heiser et al., using a computer simulation of paranoia, PARRY, psychiatrists were asked to distinguish the programme from a real patient [15]. The two hidden entities, PARRY and patient, were questioned one at a time (p. 150). The result in Heiser et al.'s study was found to be random: two of the five psychiatrists in the test thought the computer was a patient, and three of the five psychiatrists thought the patient was a computer (p. 153).

On who should question the machine, Turing had advocated an "average interrogator" in 1950. In 1952 he spoke of a 'jury-service' who should not be machine experts, and that "a considerable proportion of a jury ... must be taken in by the pretence" in order for the machine to pass the test (p. 495). Turing did introduce a *control test* in 1952, as he did in the 1950 version, with a human foil acting as witness on occasion. In the two-participant version of the imitation game Turing suggested: "We had better suppose that each jury has to judge quite a number of times", and sometimes the interrogators should face a hidden human to question, preventing them judging their interlocutor a machine in every instance of interaction "without proper consideration" (ibid).

Turing did not mention any particular length of time for interrogator questioning in 1952, unlike the "after five minutes" duration stated in his 1950 paper. However, inscribed in the 1952 discussion are Turing's eight criteria for staging the 2-participant, interrogator-witness test for *machine thinking*:

1) the machine is hidden from view and hearing from a panel of interrogators
2) the panel must not be machine experts
3) each member of the jury panel interacts one-to-one with the machine under test
4) the interrogators can only pose typed questions
5) the interrogators can ask any questions,
6) the machine attempts to respond in a human-like way
7) sometimes the interrogator is faced with a hidden human to question
8) each panel member interrogates a number of times.

Turing's two implementations for his imitation game are contrasted in Table 1.

TIG Feature	**1952 & 1950 Viva voce**	**1950 Simultaneous-comparison**
Mode of questioning	One-to-one: human interrogator-machine	One-to-two: human interrogator-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	Panel of juries	Unspecified
Number of Tests	Judge quite a number of times	Unspecified
Language for communication (e.g. English)	Same for both interlocutors	Same for all three participants
Criteria for Test Pass: Satisfactory and sustained answers	Considerable portion of jury taken in by pretence.	"average interrogator will not have more than 70 per cent chance of making the right identification"

Table 1. Contrasting Turing's two tests for machine thinking

Turing discussed educating the machine in 1948 [12], and in the first of two radio broadcasts in 1951 [17]. Turing felt that by "applying appropriate interference", the kind of interference a teacher provides to a pupil, "mimicking education" could modify a machine "until it could be relied on to produce definite reactions to certain commands" [12: p. 422]. He wrote that educating the machine requires a "highly competent schoolmaster" who must be unaware of its inner workings (p. 473), and who should transform the machine from a simple to a more elaborate system. With the aid of a *mechanic* "permitted to keep the machine in running order" (ibid), the education process

would produce a "reasonably intelligent machine" according to Turing. As he had done in 1948, in 1951 he used the example of the machine playing chess.

Turing proposed that a machine, which understood English, should be allowed to use text-based communication to record and receive remarks "owing to its having no hands or feet" (ibid). In an advanced stage of its education, the machine could forge new processes itself, resulting in highly sophisticated and highly satisfactory form of rule. Turing analogised this with engineering problems that are sometimes solved by the crudest rule of thumb procedure dealing with the most superficial aspects of the problem. Of the machine's capacity to learn *new* methods and techniques Turing suggested, "The machine's tutor ... a human mathematician ... can just present the machine with a better method whenever the machine produces an incorrect answer to the problem" (p. 470).

In addition to evolving his imitation game from a man/woman contest to a machine/(hu)man test, Turing's predictions about when a machine would pass it evolved between 1950 and 1952. In 1950 he wrote: " I believe that in about fifty year's 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" (p. 442). In 1951, in *Can Digital Computers Think?* [18] Turing wrote: "I think it is probable … that at the end of the century it will be possible to programme a machine to answer questions in such a way that it will be extremely difficult to guess whether the answers are being given by a man or by the machine" (p. 484), and "it will almost certainly be within the next millennium" (p. 486). In 1952, in the BBC radio broadcast of *Can Automatic Calculating Machines Be Said To Think?* [16] Turing responded to Max Newman's question, "But that will be a long time from now, if the machine is to stand any chance [in the imitation game] with no questions barred?" with "Oh yes, at least 100 years I should say" (p. 495). Turing really expected a machine would be considered thinking in the 21st century, after successfully engaging in question-answer tests producing satisfactory and sustained responses.

3 HAYES AND FORD'S CRITIQUE

By 1995, because no machine had succeeded in 'Turing success', that is, no machine had achieved a 30% per cent deception rate in unrestricted question-answer tests, Turing's ideas became onerous for the new science of artificial intelligence [1]. Hayes and Ford even proposed Turing's imitation game should be moved from text books to history books (p. 972). Their argument was based on the performance of machines in the first Loebner Prize for Artificial Intelligence[1] staged in 1991 [19].

In a critical article, Hayes and Ford wrote that Turing's vision was "actively harmful" to AI (p. 972 & p. 976), and that it was damaging to the public reputation of the science (bid). They also felt that the imitation game was "too closely bound up with natural language understanding to now be a beacon for the entire field" (p. 973). They further claimed that "AI is the proud heir of Boole, Babbage, and Turing, but not of Mary Shelley {Frankenstein story}" (p. 976). Hayes and Ford contend that:

i) the [Turing] test fosters the development of a "mechanical transvestite" and an "artificial con artist" (p. 973)
ii) using the imitation game "to define AI, even loosely, leads to the field disowning and rejecting its own successes" (p. 974)
iii) success (of AI systems) "sicklied over with the pale cast of Turing test insufficiency" (p. 975)
iv) (AI) "fuelling technical revolutions and changing the world, but Turing's ghost orders (AI practitioners) to disinherit successes" (p. 975)

But Turing did not propose building a mechanical con artist, nor did he guide, as Hayes and Ford seem to believe, that a "good imitation-game judge" should know all about womanhood (p. 973). This would not be difficult when the judge *is* a woman. Additionally, it is not Turing's ghost which disdains over whatever is accepted as 'AI success', it is those inside and outside the science who may expect too much from imitation-game-passing systems. Hayes and Ford believe it to be "too difficult to obtain unbiased judges" in order to interrogate the machine (p. 974). However, this has not been found to be the case; only one judge-interrogator appeared to act in a way that may be seen as biased in a Turing test contest[2]. This occurred in the 13th Loebner Prize (see footnote 2). Hayes and Ford misunderstand 'Turing indistinguishability' by claiming "one of the players must be judged to be a machine" (p. 974). This is not what Turing stated; an interrogator questioning two hidden entities in parallel may judge both their responses to unrestricted questions as satisfactory and sustained. Hence, in this situation, an interrogator can rank both hidden entities as human – if one turns out to be a machine then the machine has deceived the interrogator. Indistinguishability ranking was allowed in Shah's 2008 study [20] practicalising the imitation game across three experiments. (NB The 2008 experiments included child judges and child/teenage human foils).

In 1948 Turing had deemed the learning of languages as the "most impressive" human activity, but he did realise that it seems to "depend rather too much on sense organs and locomotion" (p. 421). Nonetheless, Turing was an optimist about what a machine would be able to do, in contrast to Hayes and Ford's view that "perhaps human conversation will always be beyond computer abilities in its complexities and subtlety" (p. 976). IBM's Watson system has shown a way forward. Its technological feat and Q/A success is discussed in the next section.

4 IBM WATSON

Turing had written in 1948 [12] that "research into intelligence of machinery will probably be very greatly concerned with *searches*" (p. 430), and that such searches would be called 'intellectual searches' (ibid). IBM's Watson system [6] is

[1] Home Page of the Loebner Prize: http://www.loebner.net/Prizef/loebner-prize.html

[2] 2003 Loebner Prize Results. One of nine judges, J9, gave nine of the ten hidden entities, including two humans, a score of "1.00=definitely a machine": http://www.loebner03.hamill.co.uk/results.html

described as a *deep* question-answer (Q/A) technology[3]. The architecture is designed to "further the science of natural language processing through advances in question and answer technology" (p. 2). Peter Norvig [21] explains that Watson's system, "relies on massive amounts of data, spread over hundreds of computers, as well as a sophisticated mechanism for combining evidence from multiple sources" (p. 4).

Watson is a "massively parallel" system which has "probabilistic evidence based architecture" that allows it to apply a hundred different techniques to "analyze natural language, identify sources, find and generate hypotheses, find and score evidence, and merge and rank hypotheses" [6: p.3]. Norvig states Watson is not connected to the Internet, that its "software uses 3,000 core computer processors taking up the space of about eight refrigerators. Its hard drives are loaded with terabytes of books and information, which Watson searches with thousands of algorithms simultaneously" (p.6). The IBM team of researchers behind Watson, naming the system after the founder of the company, worked to develop a question-answer system for four years, because companies increasingly capture "critical business information in natural language documentation" (p.2). Thus, there was a need to optimise analysis of the content of information held to answer questions with precision (ibid). Norvig adds that "one of the biggest issues is determining what the question is really asking, translating a 'natural language' query into something Watson can understand and find the appropriate answer" (p.6). The IBM team feel that advances in Q/A technology could help support professionals in critical and timely decision-making, in areas including knowledge-discovery.

In contrast to Eliza's (approximately) 200 categories of knowledge [22], Watson's content is roughly equivalent to a million books [6: p.4]. Eight universities assisted the IBM team with advancing Watson's Q/A technology, including MIT[4]. The IBM team approached the US *Jeopardy!* show [7, 23], a general knowledge quiz programme that has been running on TV since 1984, because they considered its question-answer format as "the next grand challenge in computing" [23]. IBM wanted Watson to compete against humans in a contest involving understanding the complexities of natural language [ibid]. Watson's Q/A system competed against the two best human *Jeopardy!* champions, Ken Jennings and Brad Rutter in an exhibition match [7].

The format of the programme normally entails three human contestants pitted against each other, "to answer rich natural language questions over a broad range of topics, with penalties for the wrong answer" [6: p.3]. Billed as a 'man vs machine' *Jeopardy!* challenge, Watson replaced one of the three humans. The two-match contest was held over three days in mid February 2011. The show involves a host verbally giving a clue from a particular category. The contestants must provide the correct question for that clue. For example, on Day 1 of the special show, the first topics or categories were 'Literary Character APB (all points bulletin)', 'Beatles People', 'Olympic Oddities', 'Name The Decade', 'Final Frontiers' and 'Alternate Meanings' [23].

Watson cannot see or hear, hence it received clues in the *Jeopardy!* show as a text file at the same moment the clues were revealed to the two human contestants, and at the same time as the host, Alex Trebek read them out [see YouTube reference 23: 4.37mins]. Watson and the two humans had to 'ring' (press a button) to respond to a clue, the fastest to ring was allowed to answer. Watson's avatar was a sphere with threads, or *thought rays* traversing it (see figure 1). The rays changed colour to show Watson was 'thinking': when Watson felt confident in an answer, its avatar turned green, when it got an answer wrong it turned orange, and the lines sped up when Watson's algorithms worked hard to find a clue [23: 5.38-5.57 mins]. During the game, Watson's avatar displayed an answer panel which showed the system generating thousands of possible answers for every clue, while the machine processed complex computations that narrowed down the possibilities (ibid).

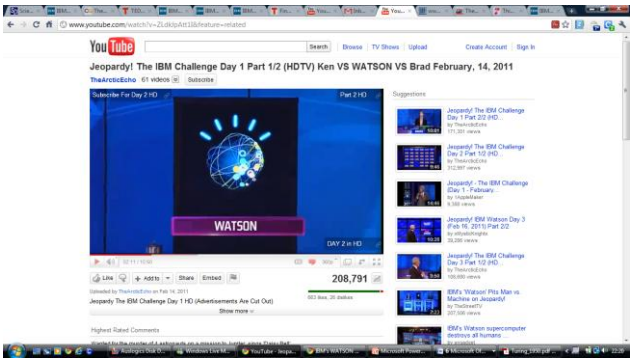


Figure 1. Avatar of Watson in the man vs machine *Jeopardy!* match

The matches involved the contestants selecting a category for a monetary reward. The first category chosen on Day 1 was 'Alternate Meanings' for $200 by contestant Brad Rutter. The first clue given in that category was "4-LETTER WORD FOR A VANTAGE POINT OR A BELIEF" [23 7.16 mins]. Rutter was the first to answer, and because he gave the correct question "What is a view?", Rutter was allowed to pick the next category. Rutter chose 'Alternate Meanings' again, but this time for $400. The clue was "4-LETTER WORD FOR THE IRON FITTING ON THE HOOF OF A HORSE OR A CARD-DEALING BOX IN A CASINO". Watson was the first to respond with the correct question "What is a shoe?" [23]. By the end of the third day, Watson had amassed a winning score of $77,147 in the two-match special quiz, compared to the human scores, Jennings on $24,000 and Rutter on $21,600[5]. In a "victory for science" [24], Watson had beaten its two human competitors to win the $1million *Jeopardy!* man vs machine grand prize [7].

Watson's *Jeopardy!* performance, *knowing what it knows and knowing what it does not know*, and being able to express this through a clue-response natural language test [23] demonstrated a level that IBM refer to as "human-expert" [6: p.5]. This was the second IBM man vs machine success. While its Deep Blue machine beat a world grandmaster, Gary Kasparov, at chess, Watson's victory in the general knowledge Q/A contest allows

[3] IBM – Watson: http://www-03.ibm.com/innovation/us/watson/

[4] IBM Press release, announcing 8 universities collaborating with Watson team: http://www-03.ibm.com/press/us/en/pressrelease/33636.wss

[5] *Jeopardy!* man vs machine scores: http://www.dailyfinance.com/story/company-news/watson-Jeopardy!-man-vs-machine-final-winner/19846648/

the IBM team to declare: "Watson's ability to understand the meaning and context of human language, and rapidly process information to find precise answers to complex questions, holds enormous potential to transform how computers can help people accomplish tasks in business and their personal lives" [6: p.6]. Perhaps it is now time for an organisation, or individual to sponsor the next and grander challenge combining Watson's impressive natural language technology with robot engineering to compete in Harnad's Total Turing Test [8].

5 CONCLUSIONS & FUTURE WORK

This paper argued that there is one Turing imitation game to assess a machine's capacity to think through satisfactory and sustained responses to unrestricted questions. The game has been misunderstood and judged according to the performance of systems in the Loebner Prize. The researcher presented Turing's own work detailing how his imitation game can be implemented in two different ways: the one-to-one interrogator-machine witness test and the simultaneous comparison of a machine with a human. Looking towards a comprehensive intelligence test for a machine, the author suggests interdisciplinary research combining the mechanical and electrical engineering of Honda's ASIMO[6] robot with the natural language power of IBM's Watson for Harnad's Total audio/visual Turing Test.

REFERENCES

[1] P. Hayes and K. Ford. Turing Test Considered Harmful. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1:972–977, August 1995.

[2] H. Shah and K. Warwick. From the Buzzing in Turing's Head to Machine Intelligence Contests. *Proceedings of the First Towards a Comprehensive Intelligence Test' workshop.* AISB Convention, De Montfort, UK March 2010, pp 12-15

[3] H. Shah and K. Warwick. Emotion in the Turing Test: A Downward Trend for Machines in Recent Loebner Prizes. Chapter in J. Vallverdú and D. Casacuberta (Eds) *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*. IGI Global: 2009: pp 325-341

[4] H. J. Levesque. Is it Enough to get the Behavior Right? *Proceedings of the 21st Joint Conference on Artificial Intelligence*. Pasadena US, July 2009, pp 1439-1444

[5] J. Weizenbaum. Eliza: A Computer Programme for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*. 9(1), January 1966,

[6] IBM Systems and Technology. Watson – a system designed for answers: the future of workload optimized systems design. IBM Whitepaper. February 2011. http://public.dhe.ibm.com/common/ssi/ecm/en/pow03061usen/POW03061USEN.PDF

[7] Jeopardy!. Watson wins 'Jeopardy!'! The IBM Challenge. http://Jeopardy!.com/news/IBMchallenge.php

[8] S. Harnad. The Turing Test is not a trick: Turing Indistinguishability is a Scientific Criterion. *SIGART Bulletin* 3(4), October 1992, pp 9-10

[9] J. Willis and A. Todorov. First Impressions: Making up your mind after 100ms exposure to a face. *Psychological Science*. 2005: 17(7), pp 592-598

[10] J.S. Albrechtsen, C.A. Meissner and K.J. Susa. Can intuition improve deception detection performance? *Journal of Experimental and Social Psychology*. 2009: 45, pp 1052-1055

[11] A.M. Turing. Lecture on the Automatic Computing Engine. 1947. In B. J. Copeland (Ed) *The Essential Turing: The ideas that gave birth to the computer age*. Clarendon Press: Oxford. 2004

[12] A.M. Turing. Intelligent Machinery. 1948. In B. J. Copeland (Ed) *The Essential Turing: The ideas that gave birth to the computer age*. Clarendon Press: Oxford. 2004

[13] H.A. Simon and A. Newell. Heuristic Problem Solving; The Next Advance in Operations Research. *Operations Research*. 1958: 6, pp 1-10

[14] A.M. Turing. Computing Machinery and Intelligence. *MIND*. October 1950: 59(236), pp 433-460

[15] J.F. Heiser, K.M. Colby, W.S. Faught and R.C. Parkison. Can Psychiatrists Distinguish a Computer Simulation of Paranoia from the Real Thing? *Journal of Psychiatric Research*. 1979: 15, pp 149-162

[16] A.M. Turing. Can Automatic Calculating Machines be said to Think? 1952. In B. J. Copeland (Ed) *The Essential Turing: The ideas that gave birth to the computer age*. Clarendon Press: Oxford. 2004

[17] A.M. Turing. Intelligent Machinery: A Heretical Theory. 1951a. In B. J. Copeland (Ed) *The Essential Turing: The ideas that gave birth to the computer age*. Clarendon Press: Oxford. 2004

[18] A.M Turing. Can Digital Computers think? 1951b. In B. J. Copeland (Ed) *The Essential Turing: The ideas that gave birth to the computer age*. Clarendon Press: Oxford. 2004

[19] S. M. Shieber. Lessons from a Restricted Turing Test. *Communications of the Association for Computing Machinery*. 1994: 37(6), pp 70-78

[20] H. Shah. Deception-detection and Machine Intelligence in Practical Turing Tests. PhD Thesis submitted to The University of Reading. October 2010

[21] P. Norvig. The Machine Age – Watson takes on humans at Jeopardy!. But how close are we to a computer that thinks? February 2011: http://www.nypost.com/p/news/opinion/opedcolumnists/the_machine_age_tM7xPAv4pI4JslK0M1JtxI/0

[22] R. Wallace. The Anatomy of A.L.I.C.E. Chapter 13 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: 2008.

[23] You Tube: Jeopardy!! The IBM Challenge Day 1 Part 1 / 2. Ken vs Watson vs Brad. 14 February 2011. http://www.youtube.com/watch?v=ZLdkJpAtt1I&feature=related.

[24] N. Deleon. The Final Night of IBM's Jeopardy! Challenge: How Did Watson Do? http://www.crunchgear.com/2011/02/16/the-final-night-of-ibms-Jeopardy!-challenge-how-did-watson-do/ 16 February 2011

[6] Asimo, Honda Humanoid Robot: http://world.honda.com/ASIMO/

Human Computer Visual Test

Yaman KAYIHAN, yamankayihan@formatd.net

Abstract

With the development of abstract art movement and with the use of computer technologies since 1980's second half, visual arts are following a different path today as they have always been, and computer has already taken a place in art as a tool.

In this paper, a test which was prepared to search how much it can be differentiated whether abstract visuals are created by computer or human, is presented. 20 visuals were prepared according to the goal of this test and with the help of a chosen software using the same parameters and colours. 10 of them were made by the "Auto-Generate" facility of the software, by computer. The remaining 10 visuals were created by me using the same parameters and colours.

In the evaluation of the answers 43 people gave to the test, it was seen that correct predictions remained in the average of 51%.

1 Introduction

The people who are not engaged in making artworks, most of the time react to abstract visuals as "I can make them as well." This only remains as a thought and generally does not turn into creation. For the ones who rarely try to make it, failure is almost certain. But, maybe for these visuals which people cannot understand or even though they cannot find anything to understand when they think that they understand them, it might be necessary to evaluate the reaction of people in another way. Maybe they think that creation of these visuals is very easy or maybe that it is something ordinary and almost everyone can make them.

If this point of view is taken to be true even to a certain extent, why can't it be possible that these visuals can be produced by computer?

20-30 Years has passed since computer was given a place in art as a tool. As computers got more powerful hardware and softwares, they started to take the place of canvases and brushes more. Beyond painting, there has remained almost no photograph which is not digital. What is more to these are, films and games.

Then, can humans be given another chance? As the creation of abstract visuals are so easy, or if people think so, then it must be the time to present visuals which are made by computer.

But how much can it be possible to differentiate the visuals which are created by computer and human with the same qualities and are similar to a certain extent?

This test, having references to the computer/human test by Alan Turing [1], aims to search the answer of the question asked above.

The organization of the paper is as: In the second section, how the visuals which were used in this test were created and in the third section, how the test was applied is explained. In the fourth section of the paper, the analysis of the data results of the test are stated. The fifth section includes the conclusion.

2 Generation of the visuals used in the test

2.1 Software

Finding the software which was used for developing this test was the most crucial step. "Alchemy version beta 008" © 2007-2010 Karl DD Willis, Initiated by Karl DD Willis and Jacob Hino was used as the software of the test [2].

The brief information about the software which is stated in its own web site is as: "Alchemy is an open drawing project aimed at exploring how we can sketch, draw, and create on computers in new ways. Alchemy isn't software for creating finished artwork, but rather a sketching environment that focuses on the absolute initial stage of the creation process. Experimental in nature, Alchemy lets you brainstorm visually to explore an expanded range of ideas and possibilities in a serendipitous way."

2.2 Visuals

10 of 20 visuals were formed by the "Auto-Generate" facility of the used software (computer), remaining 10 was created by me with the same software. All other circumstances were the same while they were numbered randomly.

Another group of visuals were created by me with the same software prior to the preparation of this test and were added to my personal visual web gallery's [3] section number 184 [4]. An example of the visuals of this section is shown in Figure 1. For the visuals of this section, all parameters and colours were used freely by me.

For the creation of the visuals which were used in the test, instead of using parameters and colours freely, fixed parameters and colour limitations were applied so that participants could be objective between computers and human.

20 visuals which were the subject of this test and the related information could be reached also from section number 184 [5].

One of the visuals which were placed in the test is shown in Figure 2. Whether this visual is created by human or computer is not stated in order not to interfere with the test.

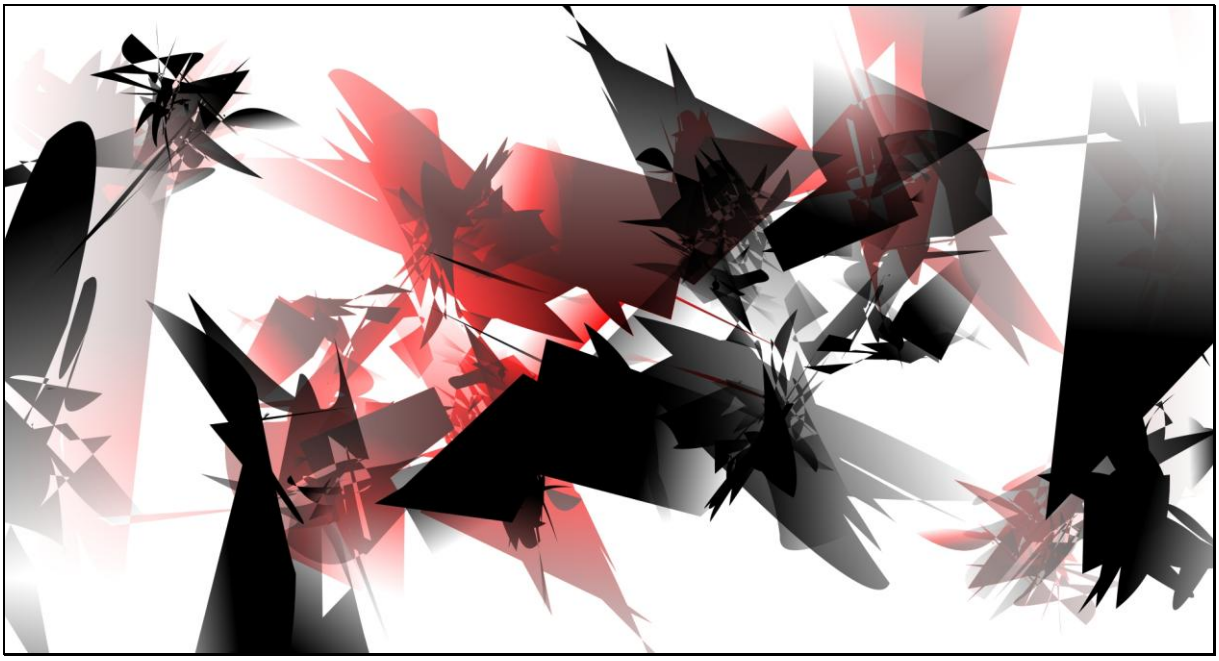


Figure 1 : One of the visuals presented in section number 184.

2.3 Colours used

In all the visuals same colours were used on the same white background in the same order and same amount (twice) :

#	Colours	HEX	Red	Green	Blue
1	Red	FF0000	255	0	0
2	Green	00FF00	0	255	0
3	Blue	0000FF	0	0	255
4	Yellow	FFFF00	255	255	0
5	Black	000000	0	0	0
6	White	FFFFFF	255	255	255

Table 1 : The colours used in the generation of visuals which were used in the test.

The colours which were used in the test were especially chosen out of the basic colours so that personal view and evaluation of the participants could be less.

2.4 Parameter adjustments used

After the choice of the software was done, the æsthetic which came to be the result of section number 184 can be said to be the basic determining factor in the preparation of this test. I used the experience I gained when preparing the visuals of this exhibition related to understanding which parameter makes what kind of effect on the result for the adjustment of the parameter values. In all the visuals, same software adjustments were used:

Parameter	Adjustment
Create	Type shapes
Affect	Gradients
Style	Solid shapes (over)
Line weight	1
Colour transparency	None
Distortion	Max
Size	Max

Table 2 : Parameter adjustments used in the generation of visuals which were used in the test.

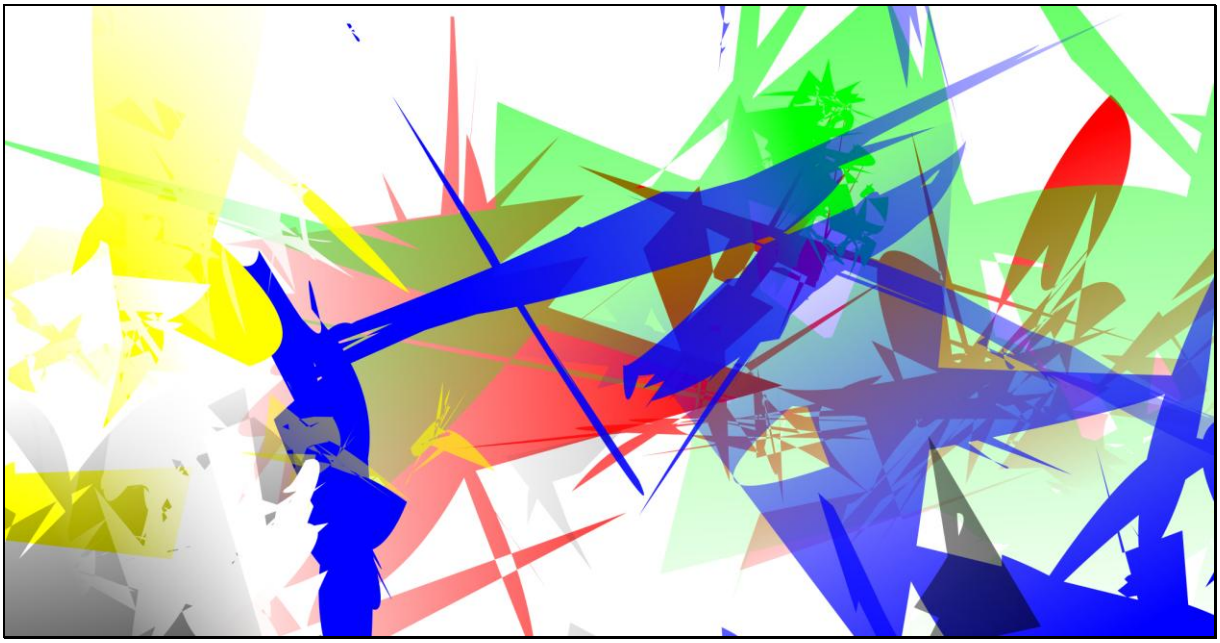


Figure 2 : One of the visuals which were presented in the test.

3 Procedure for the application of the test and data collection

Participants of the test were primarily asked to look at these 20 visuals presented in the web site and afterwards asked to guess which of the visuals were formed by computer, or created by human. In order to collect data, a form was attached to the web page [6]. After the form was downloaded and filled in, it had to be sent to me by e-mail.

4 Analysis of Results

Participants

43 Turkish people (21 female and 22 male) participated by December 27, 2010. The ages changed between 18 and 70 (average age = 28, SD=13). Min 2 (10%) and max 20 (100%) correct answers were given. Average success of 43 participants was 51%. There was no participation criterion; whoever aware of the test was free to participate.

Results

The data collected from participants to examine the success in the identification of visuals automatically generated (=b) and created by a human (=i). The data collected form a normal distribution according to the tests of normality ($p > .05$), which shows the soundness of the population participated in the task. Normal distributions are preferred for statistical analysis as many of the tests has the initial assumption of a normally distributed data. When the data form a non-normal distribution, it is forced to shape a normal distribution by applying log or z transformation. If the normal distribution cannot be obtained even via transformations, then, non-parametric tests are used for statistical analysis. In addition, the two groups of the visuals do not have different variances according to the test of homogeneity of variances, Levene's test ($p > .05$), which is another assumption of many statistical test such as t-test, ANOVA.

The characteristics of distributions are given in Figure 3. The quartiles, smallest and largest observations, and medians of the distributions are given in the box plots presented in Figure

3. There are no outliers in the distributions, which mean that none of the participants performed in a different way form the others.

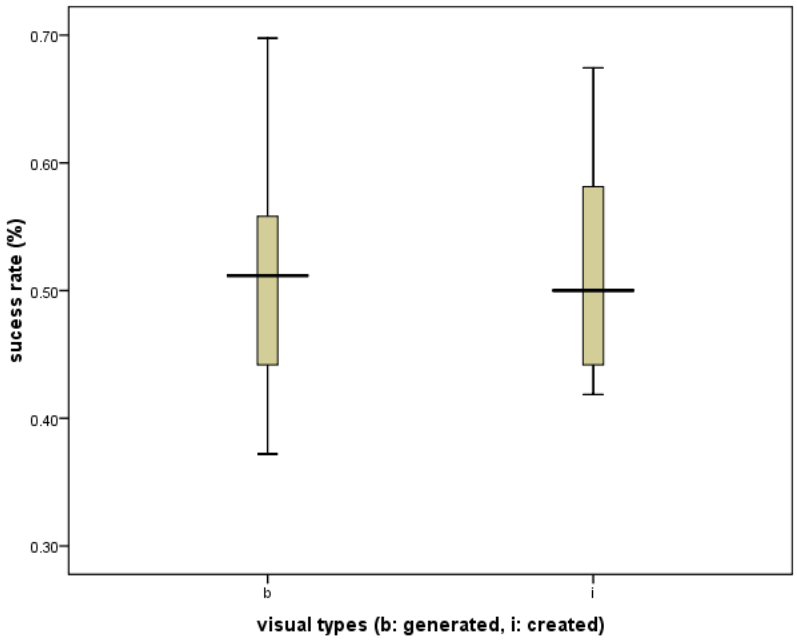


Figure 3 : The characteristics of distributions.

The success in the identification of the visuals automatically generated and created by me was examined with an independent t-test. Accordingly, there is no significant difference between the identification rates of automatically generated (M = .51, SE = .03) and created (M = .52, SE = .03) visuals (t(18) = -.23, p > .05, r = .05).

5 Conclusions

Half of the 20 visuals of the test were made by computer and remaining rest of them was created by me. This division in half might have affected the result of the test. However, that the average of correct predictions were 51% might be surprising. By 27TH December 2010, the average of correct predictions of the 43 participants of the test was 51%. The result was almost in half. So, it might be possible to say that people were neutral in the dilemma of deciding between computer and human. It might be hard to argue that the visuals created by me are artworks; however, they were at least created by a human, even though computer was used as a tool in their creation. If it is possible to say that art is an interest of humans, it might not be possible to say positive things for the part of art or human when evaluating the result of the test. However, it is easier to say positive things for the part of computer. Of course, it must not be forgotten that the visuals which were created by computer were also a result of computer and software produced by human and parameters and colours arranged by human as well. But, when the subject is the Turing test, it is certain that computer and softwares will be present.

According to my point of view, art is formed by things in which most ideas are included or maybe in this respect, art is a way of thinking. For the people who create artworks, art might be something that they live in or something that they cannot get out of. Beyond the use of computer as a tool in art, it is –at least for the time being- impossible for computers to make artworks. That it was not understood whether the visuals were created by human or computer might be related to perception. This might not have necessarily changed the meaning of art. But with this test, computer might have won a place in art beyond just being a tool.

Acknowledgements

This test and further related things are all dedicated to my identical twin daughters, whose appearances in my opinion, might be good subject to similar tests. They are my all time inspirations.

If Prof. Ugur Halici (Department of Electrical and Electronics Engineering, Middle East Technical University, Ankara, Turkiye) did not support me during the test, I think it would have never realized. I want to express my deepest thanks to Ugur Halici.

It might not have been possible to do this test without the software which was used in the preparation of the test; Alchemy version beta 008 [© 2007-2010 Karl DD Willis, Initiated by Karl DD Willis and Jacob Hino]. I thank Karl DD Willis and Jacob Hino.

I thank PhD student Gulsen Yildirim (Informatics Institute, Middle East Technical University, Ankara, Turkiye) for the statistical analysis of the test and for her contribution in evaluating the results of the test.

I also thank to my daughter Cigdem Kayihan (School of Foreign Languages, Hacettepe University, Ankara, Turkiye) for the English translation and interpretation of this paper.

References

[1] -, "Alan Turing", Wikipedia, Web, http://en.wikipedia.org/wiki/Alan_Turing

[2] Karl DD WILLIS and Jacob HINO (Initiated by), "Alchemy",
Alchemy version beta 008 © 2007-2010 Karl DD WILLIS, Web, http://al.chemy.org/

[3] Yaman KAYIHAN, personal visual web gallery,
http://www.formatd.net/yaman/kayihan.html

[4] Yaman KAYIHAN, visual generated using Alchemy,
http://www.formatd.net/yaman/SERGI184/index184.htm

[5] Yaman KAYIHAN, "Human Computer Visual Test",
http://www.formatd.net/yaman/SERGI184/test/index-test.htm

[6] Yaman KAYIHAN, "Human Computer Visual Test" data collection form,
http://www.formatd.net/yaman/SERGI184/test/test-form.doc

Yaman KAYIHAN in brief

He was born in 1956 in Ankara, Turkiye. He works for an international company. He has twin daughters.

He started painting when he was in the primary school. He opened his first 2 personal painting exhibitions in his high school years, and the other 2 in his university years. He painted works with designs/patterns specific to himself. In the first years, he used gouache, pastel crayons and oil paints. Later, he made his paintings by using computer as a tool. Recently he has been interested also in photography. Despite his use of photography as photograph, he sees photography more as form and colour, and he tends to experiment in photography. His visuals have been presented in the internet for 10 years. The first painting he made with his own patterns dates back to 1971 and for this reason he is celebrating his 40^{TH} year in 2011.

He founded the visual arts group named ">format D" (http://www.formatd.net/index.html) in 2004 with Ugur HALICI. 77 screenings were prepared with the visuals made by the members of the group. In these screenings, 31,346 visuals were presented and the total length of them was 1,277 minutes (2010). These screenings were presented many times in Turkiye and other countries.

His visuals, either drawing or photograph, are placed in his Web site in the address: http://www.formatd.net/yaman/kayihan.html. In his personal Web site, he has 9,253 visual works in 187 sections (2010).

Contact

Yaman KAYIHAN
Mithatpasa Caddesi 13
Ankara 06420 Turkiye
Phone 903124318252
yamankayihan@formatd.net

Reference Object Selection Intelligence (ROSI) Test

Antony Galton, Ed Keedwell, and Mike Barclay
College of Engineering, Mathematics, and Physical Sciences, University of Exeter

Introduction

Our proposed competition is to produce a machine which can emulate human performance in the task of selecting reference objects for answering spatial location queries given a 3-dimensional scene containing a number of objects. An example of such a query might be "Where is the red book?", to which appropriate answers might be "On the table" or "Next to the green book", whereas other answers such as "Above the door" or "In the room" may be less helpful, and therefore less appropriate. Previous work on this problem with human subjects has shown that this is something which humans can generally perform with little difficulty, although rather little is known about the factors which are involved in making the selection; for further details, see [1, 2, 3]. In particular, it should be stressed that there is no one correct answer, though some answers are clearly better than others. Note that in the above work, the prime focus of interest is in the selection of the reference object (the table, the green book, and the door in the examples above), not the choice of preposition ("on", "next to", "above", or "in"), since the latter is to a considerable extent conditioned by the former. It should be emphasised that the reference object selection task is quite distinct from that of generating referring expressions, as in the GIVE challenge [4]; at present, the corpus for the latter is unsuitable for our proposed competition, lacking sufficient diversity and number of objects in each scene.

The ROSI contest tests a small but fundamental element of human intelligence that removes the requirement for sophisticated natural language processing techniques. The ability to reference objects in a spatial context underpins our ability to convey information about the world to others and thus a machine that is capable of performing well in this context could be considered intelligent in this restricted domain.

Task

The scenario for the task is as follows. Given a scene populated by a number of different objects of various kinds, one of which is specified as the 'target' object, we have to choose some other object in the scene as a spatial reference with respect to which we may efficiently communicate the position of the target object. The objects in the scenes will be tagged with descriptive identifiers, so object recognition does not form part of the task. In comparing human and machine performance, humans will access the scenes visually whereas the machine will have access to the three-dimensional coordinates specified in the OpenGL data-files. The unit of testing ("test unit") is a scene with one object selected as target. A given scene can furnish more than one test unit, since different objects in the scene can be selected as targets.

Contest

Our proposal for the competition is thus as follows:

1. Competitors will be issued with a selection of the available test units which they may use in whatever way they wish in developing their systems (e.g., they might want to use some of them as training examples, others as testing examples). The human performance data for each of these test units will also be supplied, in the form of a frequency profile recording how often each object in the scene was chosen by human subjects as the reference object for that test unit.

2. When the entries have been received, they will be tested for their performance on (some or all) of the remaining test units, which they will have had no previous exposure to.

3. The performance of the machine will be assessed by comparing it with human performance on the same test units. Various scoring systems might be used: a simple system would be to assign for each test unit the frequency with which the reference object chosen by the machine was chosen by the human subjects, the total score being accumulated over all the test units used in the competition. The performance of any one human could also be assessed in this way, thus enabling a direct comparison between human and machine performance.

4. A machine may be judged to have passed the test if its performance is at least as good as that of the human with the lowest score (i.e., the one whose performance differs most from that of the group as a whole)

Feasibility

The test requires:

- A corpus of scenes
- Data on the human capability to process such scenes
- A standard PC on which to collect the above data and test the machine entries

A corpus of scenes: Michael Barclay has already generated a suitable scene corpus as part of his research into the reference

selection problem. An account of the early phase of developing this corpus was presented at the Aberdeen AISB workshop in 2008 [2]. The corpus consists of 133 scenes constructed in three dimensions using OpenGL. Each scene contains between 10 and 42 distinct objects (mean = 27), each of which is tagged with a descriptive identifier (e.g., "book", "table"). The scenes cover a range of types of situation, including both indoor and outdoor scenes on a variety of different scales. Two representative examples are illustrated below.

Human data: In order to assess machine performance in relation to that of humans, it will be necessary to generate sufficient data regarding the latter. We already possess a certain amount of this, gathered from human subjects who were asked to select reference objects for a subset of the test units (altogether we have 60 test units each with 20 human responses). In line with the observation above, that there is no one correct answer to any test unit, it was found that even this small group of humans were typically selecting four or five different reference objects between them for each test unit. With a larger quantity of human data, we can build up, for each test unit, a profile indicating the frequency with which each object in the scene is chosen as reference object (this will probably be zero for those objects — usually a majority — which are obviously inappropriate reference choices).

Required Funding

In order to be able to run this competition it will be necessary to do the following, for which funding is requested:

- The collection of more human data
- The development of a protocol or API for scenes to standardise machine entries
- The development of a website to publicise the contest and post "training" scenes for the use of potential entrants

Further funding could also to further research the extensions described below.

Extensions

The test could possibly be extended in the following ways, increasing complexity and introducing a greater number of facets of intelligent behaviour.

The introduction of object recognition: In the standard test, objects are labelled and so the machine is not required to recognise the objects individually. By removing these labels, a machine would need to be able to associate a 3-dimensional representation of an object with its textual description before being able to complete the object reference task above.

The introduction of object manipulation: A more ambitious extension, this would require the competitor to interact with the scene and respond to simple commands expressed in natural language such as "Pick up the ball next to the chair and put it in the box under the table". This requires the system to identify a target given its relation to a reference object, a complementary problem to that of selecting a reference object to facilitate identification of the target. Recent advances in technology provide us with a common interface to be used by human and machine candidates, namely the Wii Remote. Humans will interact with the system using the remote and the machine will have access to the same functionality via the Wii Remote API, thereby providing both systems with the same manipulation interface to the scene.

References

[1] M. Barclay and A. P. Galton. An influence model for reference object selection in spatially locative phrases. In C. Freksa, N. S. Newcombe, Peter Gärdenfors, and Stefan Wölfl, editors, *Spatial Cognition VI: Learning, Reasoning, and Talking about Space*, pages 216–232. Springer, 2008.

[2] M. Barclay and A. P. Galton. A scene corpus for training and testing spatial communications. In *Proceedings of the AISB 2008 Convention (Communication, Interaction, and Social Intelligence), April 1-4, 2008, Aberdeen, Scotland*, 2008.

[3] Michael Barclay. *Reference Object Choice in Spatial Language: Machine and Human Models*. PhD thesis, University of Exeter, 2011.

[4] GIVE: Generating instructions in virtual environments. http://www.give-challenge.org/research/. (Accessed 11/2/11).

Can Machines Think? A Proposal for an Augmented Scientific Turing Test

Patrick Fogarty[1]

"If not Turing's Test Then What?" Cohen[6]

Abstract. Motivated by French's description of the limitations of the Turing Test [8] and as an answer to Paul Cohen's question "If not Turing's Test Then What?" [6] this paper supports the development of an Augmented Scientific Turing Test (ASTT) as an inductive empirical indication of machine intelligence. It is expected that such a test would take into account changes in computer technology since Turing first suggested his imitation game in 1950 [18] and use fundamental research into linguistic development in children, intelligence measures for animals, semantic analysis and aptitude testing as a springboard for testing machines. The proposed test, while retaining the blind interrogation and opinion based evaluation of Turings's original "imitation game" see figure 1, will give an empirical set of investigations that will further probe the intelligence and responsiveness of the machines under evaluation. To provide context to the argument for such an augmented scientific test, a brief philosophical background to the attempts to develop intelligent machines will be detailed. Turing's imitation game will be explained along with an explanation of why Turing's idea, when strictly interpreted, falls short and is insufficient in a modern context. Then some kinds of intelligent behaviour that might be testable will be suggested. Considering the present day world and allowing for a broader interpretation of Turing's suggestion, the augmented scientific test will be proposed. Some arguments against the proposed test will be considered along with alternative suggestions from the literature. In conclusion it will be said why the ASTT appears to be the strongest alternative among suggested tests for machine thought.

1 Introduction

It was just after the invention of the digital computer when Turing mused whether machines could think. The idea of computing machines capable of thinking goes back to philosophical roots in the writings of Pascal and Leibniz [3, p119] and to their practical implementations of calculating machines. Many philosophers and mathematicians form the background to modern computing and it is their contributions that lead to Turing's question: Can Machines Think? [18].

1.1 Philosophical Background

Frege was the father of modern logic. His *Begriffscrift* [19], an attempt to achieve a perfect language as proposed by Leibniz [3, p119-120], saw the formal exposition of logic for the first time. It was the work of Frege and George Boole [4] that lead to the ability to produce algorithms processable by modern digital machines. Russell and Whitehead took up Frege's challenge and began to work on their *Principia Mathematica* which sought to reduce mathematics to logic [20]. However, Russell found his famous paradox that showed that mathematics could not be reduced to logic as Frege had wished [19, p124-128]. To circumvent the paradox Hilbert suggested that by using formal techniques the foundations of mathematics could be put on a firm footing by following a finitist programme [2, p183]. The work on Hilbert's programme allowed G*ö*del show how efforts to found mathematics on arithmetic are limited inherently by the fundamental properties of the systems under study [9]. However, during this work G*ö*del also proposed a numbering schema that was to be amended, adapted and combined with Boolean algebra to give rise to digital computing as we know it. Alan Turing showed that it was possible to conceive of a universal computing machine. The so called 'Turing Machine', although a theoretical machine, is the model used for all modern computers and it can be said that all computers are equivalent to a 'Turing Machine'. He also showed that such machines were limited by G*ö*del's incompleteness theorem, that is to say, that not all problems are computable. Turing did this by describing the 'halting problem' [17]. It was a combination of Turing's idea of the Turing machine and its limitations that lead Turing to ask if machines can think? These four interwoven strands lead us to modern computing and the study of AI and it's limitations:

1. The idea that laws of thought can be expressed in a binary way [4]
2. The idea that mathematics can be reduced to logical calculations. [19]
3. The idea that numerical coding can be applied to abstract mathematical symbols [9]
4. The idea that all such calculations can be done by a universal machine [17]

And so, it was the tension between the obvious success of the early computers and the limitative results that the halting problem presents that prompts the question: is the human mind limited in the same way as a computing machine? Or put simply is there an equivalence between a machine and the human mind? Can machines think? Turing asked this question and himself anticipated that the answer was positive, but in true scientific fashion he sought a test that we could use to prove that the machine was a thinking entity.

2 Turing's Suggestions

Turing suggested the question 'can machines think?'[18]. To answer this Turing did not want to become tied into endless discussion about what *thinking* is but rather decided that if the machine could copy or emulate the role of a human being in a social context then it would be thinking. To this end Turing suggested the imitation game [18].

[1] University of Sussex, email: P.Fogarty@Sussex.ac.uk

2.1 The Imitation Game

The Turing Test (TT) was introduced to the world by Alan Turing in 1950 through his seminal article "Computing Machinery and Intelligence" in the philosophical journal MIND [18]. Needless to say Turing did not call it the "Turing Test" but instead introduced it as what he called the 'imitation game' (see figure 1). The imitation game was the original variation of the test and it was defined as follows: Three protagonists, a man A, a woman B and an interrogator C of either gender, are involved. Both A and B are hidden from C with whom they communicate by means of a teletype machine. The interrogator knows there is a man and a woman but does not know which is which. It is the goal of both A and B to convince the interrogator that they are the woman.

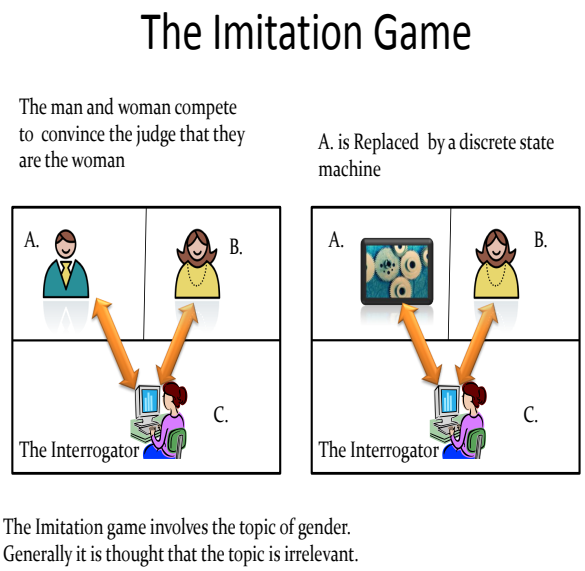


Figure 1. The Imitation Game

The Turing test as we know it has the woman replaced by a machine and the interrogator must determine which is the person and which the machine. Gender has become irrelevant in the new context and questioning does not revolve around strictly defined subject matter (see figure 2).

2.2 The Qualities of a Truly Intelligent Machine

The qualities of truly intelligent beings include

- Awareness of the environment
- Self awareness
- The ability to learn
- Tool use
- The ability to communicate
- The ability to understand

For a test to be successful in measuring machine intelligence it must address all of these items. Any single test would simply not suffice to report on all of these things. Therefore it is suggested that a battery of tests be designed to cover all that an intelligent being should be.

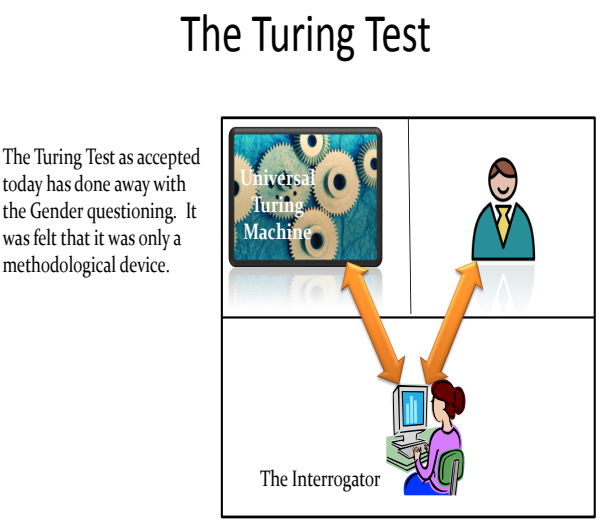


Figure 2. The Turing Test

2.3 How Might we Measure such Qualities if not Using the Turing Test?

2.3.1 A Thought Experiment:

Suppose we capture a scout from another planet and the scout manages to get away but a machine that it is carrying falls into our hands. The machine looks pretty much like a computer so we isolate it and decide to interrogate it to determine what information it can give us about the alien beings and their intentions. We must decide how intelligent this machine is since that will determine how we treat the answers it gives to any questions we pose. We in fact have to determine the following:

- Is it conscious of its surroundings?
- Is it self aware?
- Is it motivated?
- Does it learn?
- Can it survive in an unstructured environment?
- What are the limits of its intelligence?

We will not seek to anthropomorphise our captor since our very existence may depend on recognising an alien form of intelligence. It is also apparent that we will not be gentle in probing the machine but will exercise it to its maximum, probing every aspect of its being and ensuring the strictest empirical measurements are kept.

2.3.2 Intelligence in the Animal Kingdom

How do we measure the intelligence of animals? With the exception of some primates and dogs we don't tend to talk to them and expect a response. Even with the primates and dogs we don't expect

them to respond in our language. Yet we expect with clever experimentation to be able to determine roughly how intelligent a given animal is and in what way it is intelligent. Can we emulate these techniques when examining our machines? Isn't it the case that the machine is similar to the alien scout machine or the animal? It is not human and will never be human. Therefore a simple behaviour test of language skills does not suffice to show thinking since one can think without speaking. That is to say that while the Turing Test is interesting it is not necessary nor sufficient to show thinking is occurring. It is not necessary since speech is not necessary for thinking. It is not sufficient since behaviour alone does not prove the ability to think. Why judge the machine by human standards? This does not mean that one should not be rigorous in testing for intelligence or that one shouldn't use language interfaces where appropriate but just that human qualities and more specifically language are not the only measure of intelligence.

If reviews of how cognition is gauged in the animal kingdom are examined then it is apparent that several behaviours are considered. Experiments testing the following have all been done [12]:

- Social cognition - What does one individual know about an other individual?
- Self awareness
- Same and Difference - Studies of understanding the concept of identity
- Abstract and imaginal representation - The capability of forming concepts

Similar batteries of tests should be designed for our machines. The reason for the tests in our case is three fold: firstly, to provide a standard to aim at for the machines, secondly, to measure empirically how much progress is being made towards the creation of intelligently behaving machines and thirdly then to allow a phenomenological assessment of any progress:- that is, to see how socially interactive the machines have become.

3 The Augmented Scientific Turing Test (ASTT)

It is proposed that any test that is to take the place of the 'Turing Test' must achieve several objectives:

- It must be empirical and give measurable results
- It must eliminate the possibility of 'cheating' and demand that the machine under test has truly intelligent elements. That is to say, that the inappropriate use of lookup tables should be discouraged, while using logic and semantics encouraged.
- It must probe the limitations of machines as shown by the limitative results of Turing and Gödel
- It must provide in-controvertible evidence of it's findings
- It must also provide for the behaviourist phenomenological analysis as originally provided by the Turing test or in other words, the opinion of the 'reasonable man'.

3.1 The Tests

The battery of suggested tests are classified as follows:

- Formal Empirical Tests
  - Linguistic Tests
  - Environmental Tests
  - Code Examination Tests
- Informal Tests (Similar to the Turing Test)
  - Semantic Tests
  - Behavioural Evaluation Test

Each type and category of test will be treated separately and suggestions as to what they should consist of made:

3.1.1 Formal Empirical Tests

- Linguistic tests: The lingusitic tests should consist of an analysis of how the machine deals with successively more complex linguistic structures success being indicated by the fact that the machine can cope with and create complex linguistic structures or statements. The measurement will be defined as the level of complexity that the machine can achieve. Levels of complexity of structures will be defined from the linguistic research literature:- for example, using the D-Level Scale of Rosenberg and Abbeduto [13] [7] or French's distinction between the cognitive and subcognitive [8]. The D-Level scale represents the level of complexity that can be dealt with by a language user. The D-Level scale is used to analyse language development and language usage in retarded adults. In doing linguistic testing the response to cognitive and subcognitive linguistic structures should also be examined [8]. This approach would find approval from Turing as he states towards the end of his 1950 paper that

  > Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain.

- Environmental Tests: Environmental tests will consist of questions asked of the machine about its environment. For example, questions regarding what the machine is, should garner answers that show the machine is self aware and can at least deliver a rudimentary explanation of its own functioning. Further questions about the environment in which the machine performs will indicate if it can communicate with other machines in a network etc. This will indicate if the machine under test is a fully functioning self aware agent
- Code Examination Tests: Code examination will immediately identify where AI techniques are being deployed to improve the logical processing required to provide a fully self aware agent. This will help to identify successful elements and those that need to be improved or that don't work to make the machine respond flexibly.

3.1.2 Informal Tests see Figure 3

- Semantic Tests: Semantic tests will allow interrogators to probe the understanding of the machine under test. Using more and more complex structures the responses can be analysed by a human interrogator. This evaluation will help by providing data which can be used to refine the responses of the machine. The refined responses then should be more human like and provide a more comfortable interface for human interaction in following generations of machine. Semantic tests will also indicate the limitations of the intelligence of the machine and show if it can truly think.
- Behavioural Evaluation Tests: The Behavioural evaluation test will be the Turing test with the added proviso that both the human and machine can use other machines to help answer questions posed. This will give an idea of how the machine operates as

a tool user and also should preclude anyone identifying the human because of their lack of machine like qualities e.g. handling large numerical calculations. Using this technique, it should be possible to identify the machine because it lacks some very human analytical aspects, while it is not possible to identify the human because they lack machine like abilities. These tests then provide what Cohen says is inductive evidence of possible thought processes in machines [6].

3.2 Suggestion for an Augmented Scientific Test

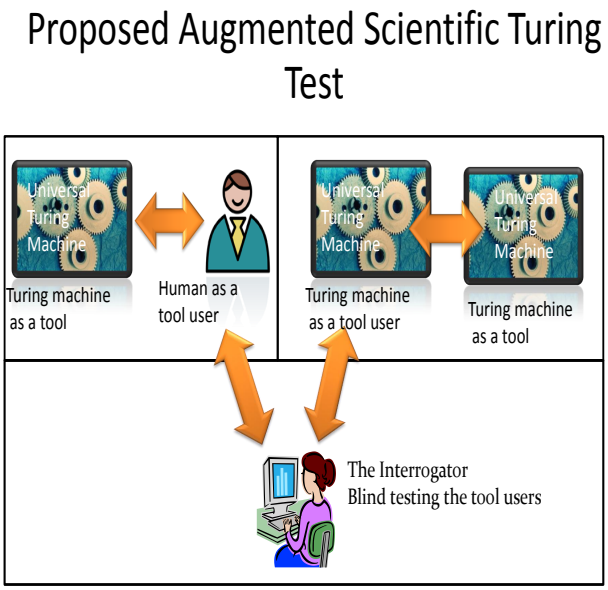


Figure 3. Turing like informal element of the Augmented Scientific Turing Test

In a contemporary context there are computing machines in every house. They can be interlinked and all have access to the shared resources of the internet. It is also clear that almost everyone has the opportunity to interact with computing machines with shared resources so that human-machine interaction has transformed society. Taking these things together it is apparent that the Turing Test is no longer sufficient to specify a test for the thinking machine. Turing's test, while containing some relevant aspects, does not represent a good empirical test of thinking or intelligence. It is based on a simple behaviourist foundation that has been shown to be lacking [16].

So to reiterate there should be two batteries of tests, the first being formal tests and the second the informal tests. Formal testing should consist of strictly empirical examinations of the linguistic abilities of the machine and an examination of it's code and inner functioning. Informal testing will examine the semantic abilities of the machine, social aspects of responses and examine behaviour as a measure as Turing had desired.

Consider figure 3, as an element in the battery of tests being suggested, the retention of Turing's social aspect is felt to be important. However, here the original Turing test has been augmented by the addition of two new computers or perhaps they could be called non-human agents or Universal Turing machines. The two new machines are added to the environment to provide the human interrogator and the machine being interrogated with tools that they can work with. This eliminates the need for the machine to play dumb by slowing or amending its responses and increases the number and type of questions that can be asked by the interrogator. In Turing's original test the computer had to wait before it could answer complicated questions such as "What is 6253547 multiplied by 353749", and now the human can use their computing tool to answer questions on complicated arithmetic and the delays deemed necessary by Turing in the machines response have been eliminated [18]. The fact that the machine's environment has now been extended means that new factors can be tested. So the machine can be tested for understanding regarding the boundaries of its capabilities. What does it know about the other individual in its environment? Can it use the tools available to it? Far from thinking it is human or pretending to be so, does the machine understand what it is? Is it self aware?

4 Possible Objections

4.1 The Separation of Cognitive and Subcognitive

French claims that it is impossible to tease apart cognitive and subcognitive questions and that therefore the Turing Test is limited since only a machine that had cultural experiences, as human beings have, could ever pass the test [8]. This in French's view make the Turing Test a test for *human intelligence* but not a test for intelligence in general. It is this observation by French that motivates this paper. However, it is felt that, not withstanding the intertwined nature of the cognitive and subcognitive, it is important to retain elements of Turing's original test. Retaining those elements allow the human machine interface to be tested and will encourage the development of more comfortable front-end access to our technologies. Essentially it is felt that French's conclusions are correct for the Turing Test but that an augmented test will address the issues.

4.2 Embodiment

It has been put forward that intelligence can only be developed in an embodied agent [15] and [10] see Cohen [6, p.486]. However, how are we to interpret 'embodied'? It is not true that the agent that resides on the internet has no environment. In fact is it not more important that an agent be aware of its environs than that it imitate being in ours? The point is that the agent should be aware of and be capable of learning from wherever it finds itself. I doubt that actual physical embodiment is a pre-requisite for thinking. There is no evidence that I can find that proves embodiment is the essence of intelligence.

4.3 Behaviourism

Behaviourists would say that the expectation that we gain anything from looking at the internal workings of the machine is misplaced. From a behaviourist point of view if the machine behaves as if it is intelligent then it simply is intelligent. It was this view that lead Turing to think of the imitation game in the first instance. However, philosophical debate since the time of Turing has tended to expect more than simple behaviourist examination, for example Searle feels that we need to look at the semantic processing of a machine in order to see that it is intelligent [16]. Therefore not to prejudge the issue the new battery of tests will cover both the behaviourist and the Searlean view and could in fact help to decide between them. So

adapting an approach that we consider neither case proved but examine the facts to see firstly can we find the distinguishing factors in the behaviourist and non-behaviourist approach and then see if results from the test help one side or other. Much is to be learned from implementing semantic processing and to encourage this it is proposed that a competition be started.

5 Competitions

There are now several competitions that draw their inspiration from the Turing Test. Below are covered some of the more significant competitions running today.

5.1 Chatter Box Challenge

"The Chatterbox Challenge (CBC) which began in 2001 is an annual contest for chatbots. It is unique in the fact that it remains a free and open contest with minimal restrictions on the type of technology used in the creation of the bot. The competition begins in mid-March and finishes by the end of April. Botmasters from across the globe submit their chatbots for evaluation and competition. Every entered chat bot is asked a series of questions and scored on its responses by a set of independent judges. The top ten bots move to a final round where an additional series of questions is posed to the finalists. The winner is selected by the judges as the chatbot who has scored the highest from among the finalists. A number of the chatbots who have entered the competition in the past have become the foundation for commercial technologies."[5] The chatterbox challenge seeks to emulate the Turing Test proper. It is run annually in pretty much the same format as the Loebner prize which pre-dates it.

5.2 The Loebner Prize

The Loebner prize is an annual prize for machines competing to beat the Turing test. It was started by Hugh Loebner in 1991 [11]. Cohen [6] believes that the prize is impotent and cannot generate interest in the AI community because of its structure and rules. It is goalless in the sense that failure does not point the way to improvement and the 'chatbots' involved are nowhere near convincing people that they are intelligent. My opinion is that Cohen is correct in his assessment of the prize and that it is time to change the format of testing and to reward innovation and progress towards truly intelligent behavior.

5.3 The Automated Negotiating Agents Competition

"The purpose of the [ANAC] competition is to steer the research in the area bilateral multi-issue closed negotiation. Closed negotiation, when opponents do not reveal their preferences to each other, is an important class of real-life negotiations. Unfortunately, the game-theoretic approaches cannot be directly applied to design efficient negotiating agents due to the lack of information about opponent. Instead, heuristic approaches are widely used to design negotiating agents." [1] This of all the competitions represents the one that could foster serious innovation and is indeed designed to do so. However, because it is restricted to negotiating agents it does not suit the domain general aims of AI agent implementations where intelligent interaction with human beings is the aim.

5.4 A proposal for a new Competition

Following Cohen [6] it is proposed that a new competition be set up not limited to the imitation game but with several categories covering:

- Awareness of the environment
- Self awareness
- Learning
- Tool use
  and special categories
- Best code and
- Best at Aptitude Tests

Such a competition should engender a spirit of adventurous experiment and lead to useful developments applicable across many disciplines.

6 Conclusions

The conclusion is that the 'Turing Test' whilst a good first effort is not fit for purpose and presents a test that encourages bad science and wasted time on research that will never be fruitful. It is time to stop fretting about the 'Turing Test' and move on with an empirical agenda to measure the intelligence or otherwise of machines using the benchmarks we happily apply to ourselves and other organisms. That is to rigorously test syntactic, semantic and behavioural ability and not just stick to the behaviourist view of Turing and Gilbert Ryle who said

> Overt intelligent performances are not clues to the workings of minds; they are those workings.[14]

A full battery of tests will ensure that we are making progress towards the behavioural aim of having socially integrated human usable machine interfaces and at the same time developing our knowledge of AI techniques, semantics and logic. It is further proposed that a new competition be instituted that tests machines for intelligence using a full battery of tests as an Augmented Scientific Turing Test. This would engender a new approach to machine intelligence and encourage good science.

And so in answer to the question 'If not Turing then what?' An Augmented Scientific Turing Test please.

REFERENCES

[1] ANAC. http://www.anac2011.com/anac-1.5.pdf, 2010.
[2] *Pilosophy of Mathematics Selected Readings*, eds., Paul Benacerraf and Hilary Putnam, Cambridge University press, 1983.
[3] Margaret Boden, *Mind as Machine: A History of Cognitive Science*, Oxford University Press, Inc., New York, NY, USA, 2008.
[4] George Boole, *An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*, MacMillan and Co., 1854.
[5] CBC. http://www.chatterboxchallenge.com/, 2010.
[6] Paul R. Cohen, 'If not turing's test, then what?', *AI Magazine*, **26**(4), 61–67, (2005).
[7] Michael A. Covington, Congzhou He, Cati Brown, Lorina Nai, and John Brown, 'How complex is that sentence? a proposed revision of the rosenberg and abbeduto d-level scale', Research Report 01, CASPR, (2006).
[8] Robert M. French, 'Subcognition and the limits of the turing test', *Mind*, **99**, 53–65, (1990).
[9] Kurt Gödel, *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*, Dover, 1992.

[10] Stevan Harnad. The annotation game: On turing (1950) on computing, machinery, and intelligence, 2006.
[11] Hugh Loebner. 2009.
[12] David Premack, 'Animal cognition', *Ann. Rev. PsychoL*, (34), 351–62, (1983).
[13] Sheldon Rosenberg and Leonard Abbeduto, 'Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults.', *Applied Psycholinguistics*, (8), 19–32, (1987).
[14] Gilbert Ryle, *The Concept of Mind*, Penguin, new ed edition edn., August 2000.
[15] Ayse Pinar Saygin, Ilyas Cicekli, and Varol Akman, 'Turing test: 50 years later', *Minds and Machines*, **10**, 2000, (2000).
[16] J. Searle, 'Is the brain's mind a computer program?', *Scientific American*, **262**(1), 20–25, (1990).
[17] A. M. Turing, 'On computable numbers, with an application to the entscheidungs problem', *Proc. London Math. Soc.*, **2**(42), 230–265, (1936).
[18] A. M. Turing, 'Computing machinery and intelligence', *Mind*, **LIX**, 433–460, (1950).
[19] Jean van Heijenoort, *From Frege to Gödel : A Source Book in Mathematical Logic, 1879-1931 (Source Books in the History of the Sciences)*, Harvard University Press, January 2002.
[20] Alfred North Whitehead and Bertrand Russell, *Principia Mathmatica to *56*, The Syndics of the Cambridge University Press, 2nd. edn., 1962.

Towards the Measurement of Plasticity and Innateness in Artificial Agents

C. White[1], **D. Bell**[1]

Abstract. A comprehensive test of intelligence must be built upon comprehensive measurement of complex contributions to intelligence. The Turing Test paradigm offers a very linear measurement and a binary output. The AI community is seeking to address this deficit and it is our position that any future test must take account of the innateness features of a natural or artificial agent and include measures to satisfy all disciplines contributing to machine intelligence. The hypothesis of the study is that significant contributions to measures of intelligence can be identified as a result of the observation that differences in capabilities of playback robots and those of more autonomous robots reflect improvements in intelligence levels. Focussing on the Innateness features included in an innovative Cognitive Architecture allows a case to be made for using readily assessable contributors to intelligence in artefacts. The potential usage of such measures is explored.

1 INTRODUCTION

In its broadest sense, Artificial Intelligence (AI) facilitates artefacts that can be thought of as being on a spectrum stretched between two end-points: task specific machines and more generic/adaptable machines (tending to autonomous) differentiated by degrees of adaptability [1]. Both of these extremes and all points in between, incorporate Innateness to some degree; in the latter case it is significantly supplemented. At this top end, roughly speaking, this spectrum corresponds to the x-axis of Fig 1. For example, It has been argued that our capacity for language is innate [2] [3]. It is important to consider the effect of innateness, which is a very subjective term, on intelligence. It is the position of this paper that, as a starting point, through consideration of traits in natural agents such as innateness, important inputs to intelligence can be made.

Suppose a test is to be set up to assess a machine's position along this spectrum. A set of grounding points to evaluate progress in intelligence at explicit staged boundaries [4] along the road from specific to general would be useful. In the engineering context of the work reported here, we refer to this as moving (e.g. 'along the road') from playback, as exemplified by say, Disneyland Robots, to 'workforward'. More broadly, with this calibration system, a shared map can be generated to align the multi-disciplined aspects of AI. Different disciplines have views on intelligence and until a unified theory of intelligence is documented, a measure cannot be devised. Therefore a comprehensive intelligence test of machines is unachievable at present. A step-wise approach to this problem is taken here.

Some of the broad current goals in AI might be summarised as follows [5]: Thinking Humanly, Acting Humanly, Thinking Rationally, and Acting Rationally. Each scenario can yield a different understanding of intelligence and consequently its measurement. However: - *"It should be meaningful in the sense that what is being measured accounts for the most general notion of intelligence"* [6].

'Intelligence testing' needs a complex approach. By common consent there are many facets to it. We focus on one of these here – the move from playback, using predominately innate knowledge (see later) to 'workforward' using this innate knowledge, supplemented and revised where necessary in the light of new knowledge acquired from experience.

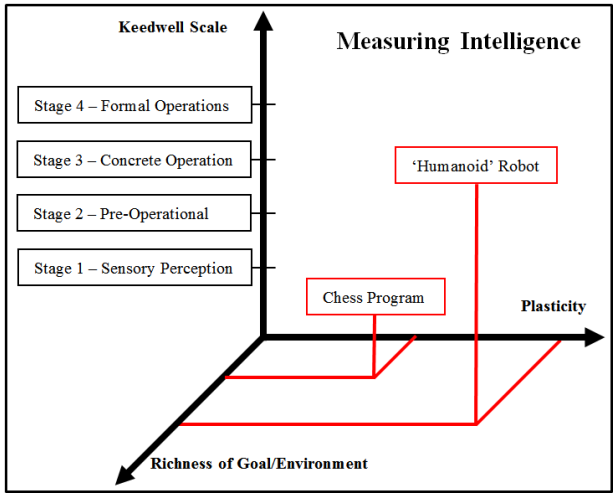


Figure 1 Dimensions of Intelligence

Keedwell [4] presents a scheme for Intelligence testing, known as the Staged Developmental Machine, based on testing methods for developing children per Piaget's theory. 'Machines could be judged by their effective stage'. It offers a scale of evaluation rather than the Yes/No outcome of the Turing test, and does not require NLP. For use in testing for intelligence levels in the present proposal, we highlight the following stages/part stages, per Keedwell, that we expect artefacts to attain:

Stage 1 Sensory Perception

- Reacts to Basic Stimuli
- 'Understands' Cause and Effect – predicts next step…
- Understands the concept of objects (those controllable and those not) and what to expect from them.
- Uses trial and error to learn about the World (experimentation).

[1] School of Electronics, Electrical Engineering and Computer Science, Queens University Belfast, 18 Malone Rd, Belfast, BT9 5BN
Email: {**cwhite06**, **da.bell**}**@qub.ac.uk**.

Stage 2 Pre-Operational

- Responding to (though NOT Language understanding) and relating to self (e.g. answering questions on its state).
- Relating objects (though NOT via language) though not in current perceptual field (memory).

Stage 3 Concrete Operation

- Conservation of volume etc. of objects(e.g. estimating quantity of objects in visual field)
- Classification of objects logical rather than on attribute basis (animals/shapes…)
- Sorting objects (e.g. size or colour)
- Effect of Reverse of action (undoing) predictable

Stage 4 Formal Operations

- Ability to create hypotheses/experiments
- Abstract thought – prediction of interactions of objects in novel ways

Weng has introduced a concept called task muddiness as a metric for higher intelligence [7]. We are concerned in this paper with relatively low levels on some of his scales. The intelligence needed for a task is measured by its aggregated multiple muddiness factors. Fig 1 is an adaptation and simplification of Weng's framework for our purposes.

Our basic approach is to take various points on the vertical scale in Fig 1**,** and seek ways of measuring the 'progress' of increasingly flexible artefacts, i.e. moving along the horizontal axis, for (initially at any rate) relatively close-to-origin points on the third axis. This paper is a basic introduction to our approach where we look at some such points.

2. COGNITIVE ARCHITECTURE

There are many Cognitive Architectures (CA) and modelling systems for cognition. The notion of having such architectures dates back to [8] where Newell identified several advantages that a unified Architecture has to offer. Examples of the advantages are: that it can help in producing a theory that is capable of helping to get a handle on the mass of data from, e.g. neuroscience studies, and that is also capable of contributing tools for addressing real-world problems. This means that some architectures have a strong commitment to have 'realistic' psychological and other 'natural intelligence' content, and others have a more pragmatic (engineering oriented) approach. The latter are not necessarily attempts at modelling the mental performance of humans or other organic agents – this represents an engineering oriented approach, and it is the one followed here.

A feature of all the architectures of interest here is that they support integrated systems of intelligent 'agential' behaviour, rather than focus on individual functions/modules. However piece-wise improvement of individual functions/modules is something we have to engage in as a means towards this end. Clearly there are many features shared by the numerous offerings in this domain. All share the basic format in the diagram below (see Fig. 2); memory, various cognitive functions, such as those for reasoning and learning, and interfaces with the outside world. In our lab we are developing a software architecture that integrates various such components of cognition in one autonomous system. In particular we focus on Innateness and Introspection.

To declare our position immediately we emphasise two architectural features in our work, both closely related to Innateness and Introspection. The first of these, which is our focus here, facilitates Adaptability/Flexibility….often achieved through some form of learning and then harmonisation, combination, and reconciliation with existing (perhaps innate) knowledge. Most CA's include modules to support the capability of an agent to acquire new knowledge from its environment through sensory-motor functions. We use the term 'learn' to refer to this. Although learning is a key characteristic of humans, it is not axiomatic that agent learning should exclusively emulate human or other higher animal learning.

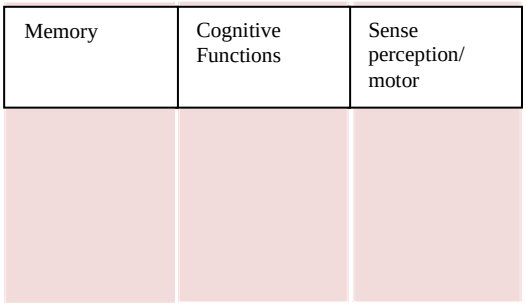


Figure 2 Basic modules of a cognitive architecture

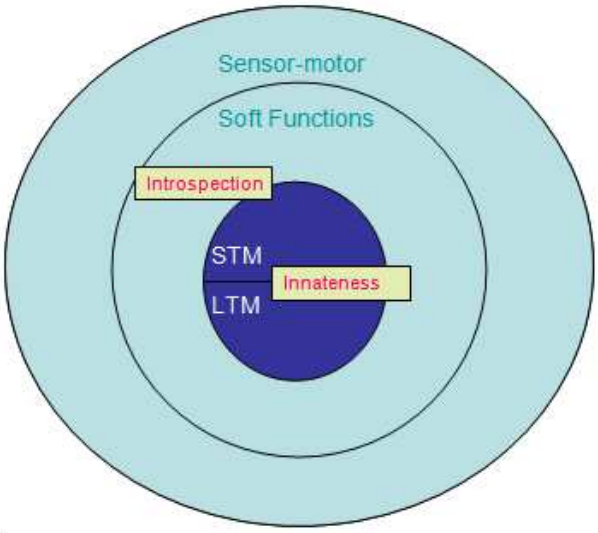


Figure 3 CA Schema

Engineers are not really concerned about how natural learning works, but they are concerned with how efficient the learning is and how effective it is seen to be after the system has learned with respect to the *ab initio* state (e.g. with respect to innate knowledge alone). Clearly this is closely related to intelligence and it something we wish to measure as a component of intelligence. For example, many robot systems are 'playback' – e.g. in Disneyland, they look intelligent in very constrained environments, but they are soon found out if they are put in different environments. Changes in environment give a way of assessing intelligence – e.g. Sphex [9] looks at sensitivity to small changes, and we are ultimately interested in how well an agent deals with complex environments. Autonomous robots require flexibility and adaptability - they are 'workforward'. Intuitively, the better they are at dealing with environmental changes, the more intelligent they are. As they work forward, changes from their innate knowledge can help in the measurement of intelligence.

The second conspicuous feature we emphasise in our architecture, but which is beyond the scope of this paper, is

'Deep Reasoning' – the ability to express knowledge and reasoning 'episodes' in terms of first principles. It focuses on the Introspection aspects of our architecture. Warrant statements, going right back to grounded concepts if necessary, for decisions or strategies resulting from episodes should be available for inspection. Trade-offs made in problem solving/conclusion reaching should be readily analysed and easily followed.

3. INNATENESS

From a strictly engineering perspective, work into innateness has been extremely limited. Interest in Artificial Intelligence grew mid-20th century, but the concept of innateness was not considered explicitly in building a life-like agent [10]. However, the idea of innateness, having been restricted to psychology and philosophy, is finding more ground in Cognitive Science, and to an extent, engineering [12]. It often appears in implicit form, mainly due to the lack of a definitive physical model [12-14].

The term 'innate' can mean many things. Often it is in reference to as intrinsic preloaded 'unlockable' content or knowledge. This knowledge is often split between nature and nurture and can be argued to come from the interaction between the two [15]. Quite often the terms innatism, nativism, innate, and innateness get interchanged. There are subtle differences in each term; however this paper takes a broad view on all as one. The following 3 areas of innateness are representative but not exhaustive:

Philosophical – Innatism is a philosophical doctrine dating from Plato. It holds that humans are born with preconceptions about their environment in the form of functionality and knowledge. The doctrine has gone through many revisions although its essence remains unchanged. It has been most notably opposed by John Locke with his famous 'tabula rasa' theory [16] - ***''Nihil est in intellectu quod non prius fuerit in sensu.' –*** *'Nothing is in the understanding that was not earlier in the senses'*. The debate within philosophy largely exists between empiricists and nativists [17]:

'empiricists favour an initial cognitive architecture that is largely content free, and in which general purpose learning mechanisms operate on the input from the senses so as to build up the contents of the mind from the cognizer's experience of the world…'

Nativists favour: *'…facilities or principles of inference that are specifically designed for the acquisition and performance of particular cognitive tasks'*

Biological – The most classic example of innateness in human biology is a non-cognitive one - that of the innate immune system [18]. The immune system has 2 parts, an innate aspect which deals with generic threats and an adaptive aspect which is more specialised and can 'learn'. Innateness in this context has also been referred to as 'that which is specified within the genome' [15]. The problem for researchers in this domain is to define where genetic data directly influences behaviour and further adding to the problem are the many interactions that take place between a gene and systems influenced by it.

Engineering – McCarthy [10] sets out a list of abilities that might be usefully innate when developing an intelligent machine. The notion of causality, objects persisting in mind without being sensed and the avoidance of dangerous situations such as heights are such examples of innate traits that are easily recognisable in nature.

Throughout history there have been, in simple terms, two schools of thought regarding Innateness; Innatism and the famous *Tabula Rasa*.

Bootstrapping [19] is an area in which an agent learns its capabilities from scratch (tabula rasa-esque) – such as sensor functions and motor functions, and maps out spatial information about its environment. There is an evident and sometimes deliberate avoidance of separating innateness and learning by researchers [19] –

"When comparing across species, it is clear that knowledge that is innate in one species is learned by individuals in another. We focus our attention on computational modelling of the learning process, and postpone the decision of where to place the evolutionary/developmental boundary."

Indeed, indicative of the problem perceived in Robotics surrounding Innateness is this particular view of Samuels [11] –

"…the term 'learning' turns out to be almost as slippery as 'innateness'"

Further to Samuels' statement, it is prudent not to disregard the possible relationship between the two areas but also to recognise that if these areas suffer from vagueness (or 'slipperiness' as Samuels puts it) then it is necessary to investigate further. At this stage it is becoming clear that innateness and learning could potentially be key aspects in understanding how to measure machine intelligence.

While empiricists such as Locke focused entirely on the role of environmental experience in a developing human, the argument here is not against this notion and strictly for innateness, but that both ideas are complementary, and they merge at the innate-developmental boundary. The precise location of this boundary is still not clear.

4. INNATENESS IN INTELLIGENCE TESTING

Consider a case where an agent has some quantity |I| of innate knowledge – like a predisposition to attack or flee from a new object as in the example below, to start with. Suppose it acquires some new knowledge N. Can we get some contribution to a measure of intelligence from |I| and |N|? We acknowledge that we have to be careful when considering the contribution of an 'innateness-overcoming' contribution to our measure of intelligence. In a straightforward playback situation, 'flee from light' is certainly pre-programmed. However merely overwhelming this initial knowledge using 'if no attack from light then attack', while giving ostensibly greater 'intelligence', could still be considered as being playback only. The agent is pre-programmed to deal with a ***predicted*** pattern of behaviour/situations encountered by the agent. Clearly the degree of movement - |change of knowledge| - must be above some threshold – perhaps to a point along the spectrum we are considering (x-axis of Fig 1) - if we are to claim genuine improvement in plasticity, and hence intelligence.

To give an increase in capability as a 'workforward' agent, we need to account for increasingly 'unexpected' (unpredicted/not

explicitly programmed for) changes in environment (or perhaps we can concentrate on the 'object' of attention such as the light). It is the degree of 'unpredictability' the agent can cope with that we want to get a handle on when assessing this aspect of intelligence. This will depend in some way on the ability of the agent to overcome, refine or add to innate capabilities or knowledge. The degree will be very high in 'humanoid' robot agents, but it might not be quite so high for a cockroach or a pigeon. In what follows, we are looking at low valued points on the horizontal axis of Fig 1, but we conjecture that incremental improvements along this axis will be important in assessing intelligence. We start this process by looking in a little detail at some of the near-origin points on this axis.

An agent's being equipped with a general rule such as the following could help facilitate some important aspects of its ability to workforward, even at the pigeon level.

If any significant change in object's behaviour or the rest of the environment, - then approach, observe (and accumulate a score).

Significance is in the eye of the beholder here – and it is closely related to risk (i.e. the product: *probability of event * penalty*). Suppose agent A is watching agent K working as an FSA (Finite State Automaton). A can find some rules of behaviour by machine learning methods. Overall we have the following pattern.

1. Start out by Approach and Observe.
2. If the behaviour of the object follows expected (programmed for) lines, then use the pre-programmed pattern to make decisions;
3. Make periodic Observations;
4. Accumulate scores as in Table 1;
5. If the 'episode is still open' (i.e. neither attack nor flee decision has been made) then. if the amount of unexpected behaviour > threshold, Approach and Observe
6. *If |'behaviour unexpected'| > threshold, then Approach and Observe and accumulate scores.*

In our scenario, we are interested in measuring, among other things, |FSA| and |'behaviour unexpected'| for any change in the environment (or, for the latter measure, change in the FSA itself). Intuitively, if this is high enough the agent could perhaps be deemed intelligent in a Turing Test scenario. A new instinct could take over, for the next episode. So this playback is a bit little less fixed than it was before the adjustment.

Plasticity and *persistence* are important in the assessment of intelligence. There is a lot of research [20] to demonstrate that there is plasticity in cognitive functioning, so tests of intelligence should consider it. Plasticity is needed to cope with environment changes and 'interruptions'. It seems to us to be self-evident that an intelligent agent must also be persistent - e.g. Persistence here is 'keep trying to get food'.

5. MEASURING THE ABILITY TO TRANSCEND INNATE BEHAVIOUR

We now look at this scenario in more detail.

5.1 Instincts in IFOMIND (case study)

In the case of agent K being implemented as IFOMIND, the FSA model is given up front (sometimes referred to informally as 'the instincts'). It includes (at least) instincts 1-3 below:

I_1 - Flee light
I_2 - Approach and Observe
I_3 – Attack if OK

The behaviour pattern can be made a little less deterministic by including some uncertainty. There are two kinds of machine learning (ML) here:

Functional ML – acquire short-term Knowledge of a particular instance for immediate use – e.g. seeking food under the possibility of danger.

Adaptable ML – motivation/purpose (e.g. hunger-inspired) can, on acquiring some sense data, *weaken* prior instinct / overcome instinctive Knowledge and change the outcome.

Other learning that is not of direct interest here also takes place on the part of K in this scenario.

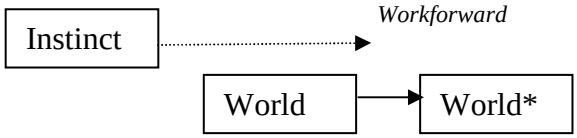
5.2 Performance Indices of Intelligence based on Instinct and Innateness

Instincts are innate – by which we mean that they are available *directly*, in a particular sense of that word. Natural lines of inheritance clearly have a big part to play in organic agents, in the existence of some habitual patterns of behaviour or thought. Such traits come, to some extent and in some way, biologically from parents. In artificial agents they are usually made available through pre-programming. For example, a vacuum cleaner could come to the purchaser as a system that can 'cover' a rectangular room; adaptability is built-in so that it can customise to a particular rectangular room, and then to rooms that have chimney breasts, etc. Just as for K above, later episodes are started with details of the particular rectangular room, less its excluded zones. So the initial system is flexible enough for this and it can be moved to another room, and adapt to it. This is more flexible than closed playback, but it can only cope with a very limited type of world, so intuitively it exhibits only a modest level of intelligence. Contrast this with a human's capability of carrying out a similar function, but also working out, for example, how to operate a new video recorder, or planning a vacation around the world, effectively concurrently.

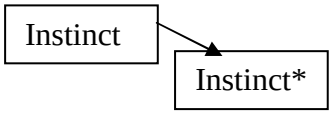
On the other hand acquisition of traits is not independent of the action of the environment – 'nature and nurture' go together. As we shall see below, instinct goes together with learning – the innate with the acquired and suitable opportunities for learning can be experienced. Traits combine inputs from instincts and inputs from experiences in an environment, and they equip an agent to deal with that particular *type of environment* – one that has 'matching properties' in some sense. In the room example, the property that all walls must be straight, and there are only rectangular 'intrusion' such as chimney-breasts. The instincts are a set of (possibly 'underdeveloped') traits that are evident in the activities of naïve agents – even before any contact with the environment. NB. 'What constitutes the minimal set of these that

constitute cognition in a particular situation?' is a very interesting research question for future work.

Learning is taken here to be the acquisition of knowledge or know-how about the world *from* the world [9]. This knowledge will often have behavioural manifestations, and after its acquisition there will be some sort of enhanced 'harmony' between agent and environment. Another question that we are interested in, beyond the scope of this paper, is: When is the agent conscious/aware of having new knowledge or, more simply, new information? Usually we assume that there are objectives (e.g. *s* and *f* in the thought experiment below), however basic, that the agent is pursuing, and the question arises as to how to represent these objectives to the agent itself. The representation should include ways of ensuring persistence and plasticity in the face of interruptions, and environmental changes (*world* changes to *world** below). The agent's ability to recover from disturbances such as these – *the plasticity/persistence index* - is a key *index of performance* [8].



When measuring intelligence we need to specify a triptych of values, viz: *{index of performance, workload – type of environment, functional features of agent}.*
An index which is related to that above (i.e. Sphex-like) is the degree of capability the agent has *to adapt by revising and updating innate knowledge.*



In particular, this distinguishes playback from autonomous (workforward) agents, as an extreme situation. In playback systems the inflexibility of instinctive behaviour is due to the lack of the properties of plasticity and persistence that we seek. The results of learning supplement the agent's innate knowledge (e.g. in humans… deontic predisposition, mathematical principles, continuity of experience). The new knowledge can be the input to the next episode, or the agent can return to start state.

A further index of performance is the degree of *improvement of harmony* with the environment mentioned above – assumed to be roughly measured by closeness to the task objectives of the agent - which the agent achieves through learning and reasoning. This is related to the *utility value index* of these cognitive processes.

5.3 Scenario for Adaptation

Notation: We have a set of Objectives $O: = \{s,f\}$ where *s* is 'stay safe', and *f* is 'get food' and a set of behavioural Instincts as given above $I_b := \{I_1\text{-}I_3\}$.

I_1 - Flee light
I_2 - Approach and observe
I_3 – Attack if OK

r_1 is the result obtained when I_1 is followed, *etc.*…

Assume that *s* outweighs *f initially*, but that this can change, for example by *habituation*. In psychology, habituation is learning in which there is a progressive change – e.g. reduction - in the probability of a (often behavioural) response as a stimulus is repeated. For example, the agent responds initially to the stimulus, e.g. to the light, by instinct I_1, but if nothing happens to affect the objectives $O = \{s,f\}$ - that is, there is neither 'reward' (Food) nor 'punishment' (not Safe) - subsequent responses are perhaps reduced in strength.

An example of the result of this in the real world can be seen in situation where a Peregrine Falcon shape has been mounted on a street-light column to try to diminish the annoying attentions of sea-gulls and pigeons to picnickers. The birds presumably react to it at first as though it were a real predator, but they typically react less as time went on, showing habituation, until some birds actually perch on the shape! Habituation has been observed in most animals. We add the following notation:

m(o) denotes 'objective *o* is to be met'
$m(o)\text{->} I_b$ denotes 'objective *o* is strong enough to make I_b dominate'
δ denotes 'required threshold strength of relevant stimulus is received'

So in our bird scenario in above..:

I_{b1} be the instinct 'to flee from a falcon'
I_{b2} be the instinct 'to scavenge from picnickers'

$m(\neg s)\text{->} I_{b1}$ denotes 'objective *s* is strong enough to make I_{b1} dominate'
$m(f)\text{->} I_{b2}$ denotes 'objective *f* is strong enough to make I_{b2} dominate'

When both objectives *s* and *f* are to be met, we take account of any ranking between them. For example, suppose the second objective overwhelms the first. We can express this as follows:
$(m(s),m(f))\text{->}_{\beta} I_{b2}$ denotes 'objectives *s,f* taken together are of strength enough to make I_2 dominate'. We allocate a strength to this requirement, in this case β.

Let us suppose the gulls and pigeons first flee the 'falcon'. We use two 'deltas' for notation… ∂ denotes 'required threshold strength of relevant stimulus is received' for the relevant Instinct to 'fire'. This is related to the calibration of 'ground points' on the x-axis of Fig 1. In the case above, in the full scenario, a small negative score comes from each lack of attack by the 'falcon' and it is accumulated until ∂ is reached. For the other 'delta', $(r_1, \delta) \text{->}_{\gamma} I_2$ denotes 'the result of following instinct I_1 with degree δ invokes I_2 ' with strength γ

6. ACCUMULATING THE δ s

As a simple exemplar we seek a measure of how much 'accumulated' δ is needed to bring about some decisive change of behaviour. This can give input to the calculation of the *plasticity/persistence index* above. It is subject to, or a function of, the richness, complexity and other features of the o*bjective*, the *complexity*, etc. of the relevant aspects of the environment, and *type of environment*, the *speed* S, or *rate of change*, at which the adaptation is made, the *size* |U| of the innate information that is used, and its *complexity*, and the size |C| of the innate information that is changed, and its complexity.

Here we consider just one aspect of this calculation.

The level of ability to overcome 'playback' in order to achieve a different final decision can be a measure of adaptability, and therefore gives a component of an intelligence measure.

We can measure empirically the degrees of inconsistency between an agent's learned model and the real world (assumed to be fully understood), and this can be compared to the

inconsistency between the instincts and the real world model. This can be repeated for different environments, and with different sizes of initial knowledge, |U|, to gain insights into the assessment of one component of intelligence.

Action/ Stimulus	*Outcome*	*Score Added*	*Result*
look/light	B+E	0	S
look/light	B+E	0	S
look/light	B+E	0	S
test/react	E+O	-0.5	S-0.5
test/react	E+O	-0.5	S-1
test/no attack	E+O	1	S
test/no attack	E+O	1	S+1
test/no attack	E+O	1	S+2
test/no attack	E+O	1	S+3
test/no attack	E+O	1	S+4

Table 1 – Trace of action: steps in a behavioural episode

Consider the trace given in Table 1. Instincts here are concerned with whether to investigate/explore (E) or Beware (B). Both of these trigger Observations (O) from which the robot learns. When a new object appears – both E and B lead to O plus learning, and this takes place systematically. The decision on whether E or B dominates at the end depends on the evidence – from sense data. S is an initial score used as a control variable.

Notice that in this kind of 'learning' we do not use the normal pattern of counting frequencies – instead we accumulate the experimental results until a threshold is exceeded.

However it is possible to get a handle on all of this theoretically as follows. At the end of the episode in Table 1, we are at the point in the trace the accumulated score exceeds the original score (S) + 3 (so B dominates here – object/light is dangerous).

Episodes are captured as behavioural traces, such as that illustrated in Table 1 – the result depends simply on the rules (instincts) and the sense data**.** At the end of this phase, when some termination criterion is reached, the balance of evidence lies in some particular direction- e.g., *the object is dangerous*. The termination criterion, ∂, for the tabled data, occurs when the accumulated 'score' adds >3 to the current score S.

Persistence is demonstrated here. In spite of some reaction from the object, the agent does not attack right away. It persists in 'hoping' to get food. The measure of persistence is related to the number of 'unsuccessful tries'-'no attacks experienced' – that are accommodated before giving up if threshold is not reached (the desired output was obtained earlier). Another input to the measure of persistence, here 'S increased by 3' , is due to the innate capability to wait for some time lag until S-> S+4, before coming to the conclusion ' object S is not dangerous'.

The plasticity here is evident in the coping with the change in 'environment 'due to an encounter with an unrecognised new object, which could either be food or be dangerous.

General Rules:
Rx: new things might be food - explore
Ry: new things might be dangerous- explore

Persistence Rules:
R1: keep observing/trying to get food
R2: keep observing/trying to be safe

Plasticity Rules:
Ra: If light detected the dangerous
Rb: If no attack after t tries, then not dangerous

Rules x and y are not revised – they remain in force to deal with further new situations. The initially-obtained knowledge that the object is dangerous (result of invoking y) is revised.

We are interested in finding the degree to which the knowledge can be overridden - towards a measure of plasticity. In this case it depends on the value of the threshold.

7. REALISING THE TEST

To create a test of intelligence that supersedes the Turing Test, in our opinion, requires a substantial effort from all disciplines interested to invest in a comprehensive model of intelligence. We believe that a realised version of this test would incorporate Innateness and Plasticity.

It is important to consider what aspects of an organism, for example, are present before an environment is experienced to help design what features can be built into an artificial agent. There is a special link between innateness and environment. James talks about humans being born with 'locks' to which the 'keys' are found in the immediate environment [21]. It can be conjectured that the process of adaptation to the environment creates this link between Innateness and environment.

In our view, developers of a Turing Test replacement should consider the following seven points:

1. A Generic Approach to inbuilt (innate) functionality in an artificial agent with careful consideration of
2. Plasticity of an agent – how initial innate structures can be reshaped with environmental experience, evolved through
3. Developmental learning to deal with certain and uncertain environments but also to invoke
4. First principles to solve problems while driven by a
5. Modular Cognitive Architecture and it should include an
6. Introspective function that can display agent understanding of its actions/environment/sense data [22] and
7. Prediction of future outcomes based on experience.

If these considerations are taken into account, then the AI community will widen the gap between task-based AI and adaptable AI and hence take a step in the direction of autonomous agents.

8. CONCLUSION

In attempting to construct a comprehensive test for machine intelligence one must consider as many facets of intelligence as possible. This conceptualisation must encompass multi-disciplinary views on intelligence and address any points of contention systematically.

The aspect of intelligence addressed in this paper is the concept of innatism from an engineering perspective. The study of innatism spans the realms of Biology, Psychology, Philosophy and Computer Science, and many commentators argue that innate traits must be built into an entity which is called 'intelligent'.

Overwhelming initial knowledge to deal with unpredicted patterns of behaviour/situations encountered by an artefact increases its plasticity and persistence – prime factors in determining the level of intelligence of the artefact. Clearly the

amount, and perhaps rate and other features, of change of knowledge must be above some threshold if we are to claim genuine plasticity, and hence intelligence.

We also need to account for increasingly 'unexpected' (unpredicted/not explicitly programmed for) changes in the environment. The degree of 'unpredictability' the agent can cope with helps when assessing this aspect of intelligence. This will depend in some way on the ability of the agent to overcome, refine or add to innate capabilities or knowledge.

Some preliminary ideas for doing this are presented in this paper, for consideration within the framework presented by the Cogio Cognitive Architecture. In this programme we will use the Webots software environment to prototype episodes where perturbations of the background knowledge and the environment of an agent can be controlled, and the effects can be determined and used to see the impact their impact on their capabilities.

ACKNOWLEDGEMENTS

This work was made possible with the funding of the Department for Employment and Learning Northern Ireland (DELNI). I would like to thank the department for their continued support throughout my research. I would also like to thank my supervisor and the Chair of the Postgraduate Research Committee for the School of Electronics, Electrical Engineering and Computer Science at Queen's University Belfast for their continued support, belief and interest in my work.

REFERENCES

[1] Peters, J.; Morimoto, J.; Tedrake, R.; Roy, N. (2009). Robot Learning, IEEE Robotics & Automation Magazine, 16, 3, pp.19-20.
[2] Fodor, Jerry A. The modularity of mind: an essay on faculty psychology. Cambridge: MIT, 1983. Print.
[3] Chomsky, Noam. Aspects of the theory of syntax. Cambridge: M.I.T. Press, 1965. Print.
[4] Keedwell E, (2010) Towards a Staged Developmental Intelligence Test for Machines, in Proc of Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 2ist Century Symposium, Ayesh, Bishop, Floridi, Warwick (eds) Apr 2010 pp28-32
[5] Russell, Stuart Jonathan, and Peter Norvig. Artificial intelligence. Third ed. Upper Saddle River (N.J.): Pearson, 2010. Print.
[6] Jose Hernandez-Orallo, David L. Dowe, Measuring universal intelligence: Towards an anytime intelligence test, Artificial Intelligence, Volume 174, Issue 18, December 2010, Pages 1508-1539
[7] Weng, J. Muddy Tasks and the Necessity of Autonomous Mental Development, in Proc. 2005 AAAI Spring Symposium Series, Developmental Robotics Symposium, Stanford University, March 21-23, 2005.
[8] Newell, A. (1990). Unified Theories of Cognition. Cambridge, MA: Harvard University Press.
[9] Espejo-Serna J.C., (2010) Connecting the dots my own way: Sphex-test and flexibility in artificial cognitive agents, in Proc of Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 2ist Century Symposium, Ayesh, Bishop, Floridi, Warwick (eds) Apr 2010 pp1-6
[10] J. McCarthy, "The well-designed child," Artif. Intell., vol. 172, pp. 2003-2014, 12, 2008.
[11] R. Samuels, "Innateness in cognitive science," Trends Cogn. Sci. (Regul. Ed.), vol. 8, pp. 136-141, 3, 2004.
[12] A. N. Meltzoff and M. K. Moore, "Imitation in Newborn Infants: Exploring the Range of Gestures Imitated and the Underlying Mechanisms," Dev. Psychol., vol. 25, pp. 954-962, 11, 1989.
[13] R. P. N. Rao, A. P. Shon and A. N. Meltzoff, "A Bayesian model of imitation in infants and robots," in In Imitation and Social Learning in Robots, Humans, and Animals, 2004, pp. 217-247.
[14] C. A. Calderon and H. Hu, "Robot imitation from human body movements," in Proceedings of the AISB05 Third International Symposium on Imitation in Animals and Artifacts, 2005.
[15] Elman, Jeffrey L. Rethinking innateness: a connectionist perspective on development. 2. print. ed. Cambridge, Mass. - London: MIT Press, 1997. Print.
[16] J. Locke, P.H. Nidditch, ed. (1690/1975), An Essay Concerning Human Understanding, Oxford University Press, Oxford.
[17] Carruthers, Peter, Stephen Laurence, and Stephen P. Stich. The innate mind. Oxford: Oxford University Press, 2005. Print.
[18] C. T. Singh and S. B. Nair. (2005, An artificial immune system for a MultiAgent robotics system. Transactions of Engineering, Computing and Technology 6pp. 308-311.
[19] B. J. Kuipers, P. Beeson, J. Modayil and J. Provost, "Bootstrap learning of foundational representations," Connect. Sci., vol. 18, pp. 145-158, 2006.
[20] Mercado, E,III Neural and Cognitive Plasticity: From Maps to Minds, Psychological Bulletin 2008, Vol. 134, No. 1, 109–137
[21] W. James, "What is an Emotion?" Mind, pp. 188-205, 1884.
[22] S. Harnad, "The symbol grounding problem," Physica D, vol. 42, pp. 335-346, 6, 1990.

Knowing me, knowing you: On the relevance of a mind reading test for general testing of intelligence

Elpida S. Tzafestas [1]

Abstract. This short article presents a discussion of the relevance of mind reading tests to general testing of intelligence and an example of mind-reader behaviours for the IPD. It is discussed how a mind reading capacity may allow intricate emotional behaviours to emerge and how these relate to a broader developmental context.

1 INTRODUCTION

The human quest for the design of an artificial brother is almost as old as humanity. In ancient Greek mythology, lame god Hephaestus had created himself artificial slaves to help him in his smithy. Other stories and fantasies abound ever since. The quest to design an artifact as intelligent as ourselves or even more intelligent was formulated more recently and is the central goal of artificial intelligence. Alan Turing is considered a precursor of modern AI by having provided a relatively formal answer to the question "Can a machine think?" in the form of a certain "test" that the machine should pass in order to be considered intelligent [1]. The test consists in a verbal interaction with a human that should be considered as natural by a third party in the sense that the observer should not be able to distinguish between the human and the machine. This procedure, labeled "imitation game" by Turing, has been generalised to encompass *any* sort of problem and *any* sort of behaviour that can be considered as human by an external observer. Turing's original claim for adequacy of this test has been heavily criticised, reformulated and defended throughout the years (for a not too recent overview see [2]).

In almost all meaningful interactions between two humans the full human potential for intentionality and consciousness is exhibited. For the purpose of the imitation game, in corresponding interactions between a human and a machine, some degree of machine intentionality or consciousness should be perceivable from the outside, by an external observer, otherwise the machine will be, somewhat fuzzily, considered as a "robot" rather than as a human replica. One such social feature characteristic of human nature and not of other primates or lower animals is mind reading, i.e. the capacity to "read" another person's mind and understand its intentions.

Mind reading is generally implicit baggage for any social activity and corresponding deficits to correct mind reading will lead to what will be externally perceived as lower social intelligence (see for example discussions on autistic intelligence, [3]). Because mind reading is considered unique to humans, it is also associated, although not necessarily causally, with other human monopolies, namely with intentionality, human-level imitation, language, empathy and more (see several chapters in [4][5]).

In the next section, we explain more in depth the connection between mind reading and intelligence and between testing for mind reading and testing for intelligence and in section 3 we give an example on the well known prisoner's dilemma. Section 4 discusses the implications of this approach and concludes.

2 FROM TURING TO MIND READING

The connection between mind reading and a test for intelligence is not as far fetched as one might initially think. Turing himself, in the original formulation of his imitation game [6, last section] wrote:

> *"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. If we are able to explain and predict its behaviour or if there seems to be little underlying plan, we have little temptation to imagine intelligence. With the same object therefore, it is possible that one man would consider it as intelligent and another would not; the second man would have found out the rules of its behaviour [...]"*

Not to be underestimated is the fact that Turing gave to this short last section the title *"Intelligence as an emotional concept"*. One can speculate that for a human observer of another human or machine acting, attribution of a degree of intelligence to it may have a number of emotional consequences, such as admiration, envy, anger with oneself, perseverence to own effort etc. From our everyday experience we know that such feelings lead almost unavoidably to alteration of one's behaviour, especially in cases where some competition is involved, implicitly or explicitly. For example, we may imagine a school child realizing that its fellow pupil solves difficult problems in arithmetic that he cannot solve. Apart from the birth of feelings such as the above, the school child may eventually devise ways to achieve the same performance by either making friends with the fellow or by plainly stealing his work. In the latter case, and provided that the child is not caught, an external observer will attribute to it the same level of mathematical intelligence as its fellow, based solely on performance.

A human interacting with a machine, for example in the context of a two-party game, will impulsively make use of such mind reading abilities, that generally give him an advantage against the machine. Furthermore, the machine may not be externally regarded as intelligent, if one cannot attribute to it *any* notion of intentionality which involves in most cases a degree of

[1] Cognitive Science Laboratory, Department of Philosophy and History of Science, Univ. of Athens, Ano Ilisia 15771, Athens, GREECE. Email: etzafestas@phs.uoa.gr.

other or self understanding. On the contrary, a human observer acquiring the feeling that he is being watched and partly understood by the machine will be ready to assign intelligence to the machine and probably become frightened and tempted to quit the interaction.

We depart from Meltzoff's "like-me hypothesis" [7] that connects imitation and mind reading by assuming that when infants see others acting similarly to how they have acted in the past, they project onto others the mental state that regularly goes with that behaviour. However, we do not claim that people ordinarily *think* that others are like them, but adopt the more modest view that what a machine can actually do is discover whether an observed entity is "like it" in the sense that it acts in the same way. To paraphrase Metzoff ([7], p. 75) intelligent artifacts may use their own intentional actions as a framework for interpreting the intentional actions of others.

3 AN EXAMPLE

We have implemented a mind reader version of Iterated Prisoner's Dilemma (IPD) players. IPD is often used as a benchmark for the study of cooperative and altruistic. In general, the cooperation problem between two (or more) agents states that each agent has a strong personal incentive to defect, while the joint best behaviour would be to cooperate. This problem is traditionally modeled as a special two-party game, the Iterated Prisoner's Dilemma (IPD).

At each cycle of a long interaction process, the agents play the Prisoner's Dilemma. Each of the two may either cooperate (C) or defect (D) and is assigned a payoff defined by table I.

AGENT	OPPONENT	PAYOFF
C	C	3 (= Reward)
C	D	0 (= Sucker)
D	C	5 (= Temptation)
D	D	1 (= Punishment)

Table 1. Iterated Prisoner's Dilemma score matrix

The first notable behaviour for the IPD designed and studied by Axelrod [8] is the Tit For Tat behaviour (TFT, in short):

- Start by cooperating,
- From there on return the opponent's previous move

This behaviour has achieved the highest scores in early tournaments and has been found to be fairly stable in ecological settings. TFT demonstrates three important properties, shared by most high scoring behaviours in IPD experiments.

- *It is good (it starts by cooperating)*
- *It is retaliating (it returns the opponent's defection)*
- *It is generous (it forgets the past if the defecting opponent cooperates again).*

In the literature we may also find stochastic strategies [9], studies in a purely evolutionary perspective ([10]), theoretical or applied biological studies ([11]) and studies of modified IPD versions ([12]). Noise is implemented as a nonzero probability that an agent's action will be switched to the opposite, i.e. from *COOPERATE* to *DEFECT* or vice versa. It has been shown that retaliating strategies such as TFT can score quite badly in the presence of noise, despite their superiority in the non-noisy domain [13][14]. This happens because even accidental defections may lead to a persistent series of mutual defections by both players, thus breaking cooperation. The usual approach is to introduce some degree of explicit generosity to account for opponent's misbehaviours or to attempt opponent modeling.

In earlier work of ours we have shown how a fundamental TFT-like behaviour (called Adaptive TFT) with the possibility to adapt itself to its opponent's friendliness may achieve very high scores with all kinds of behaviour, including suspicious, random and periodic behaviours [15]. We have also shown [16] how an additional mechanism of attraction may induce highly cooperative behaviour even if one of the agents is spiteful or in the presence of noise. The difficulty of tackling an arbitrary opponent of unknown behaviour, especially in the presence of noise, is to understand whether perceived defections of his are intentional or a result of mis-perception or inertia. As a simple example, two Suspicious TFT (STFT) agents, that are TFTs that initially defect, are unable to converge to mutual cooperation. An Adaptive TFT agent can solve this problem against STFT but not against another Adaptive TFT which is defective at the moment due to prolonged unhappy interactions.

A mind reader version of an arbitrary behaviour for the IPD is simply a behaviour that continuously examines whether its opponent uses the same behavioural model as itself. This is implemented as follows:

```
t = actual time (IPD game round);
for i=0 to min(t-1,w)
{
     theT = t-i-1;
     if (simHist[i] == oppHist[theT]))
            opp_like_me ++;
}
if (opp_like_me >= T) COOPERATE;
else generate_behavior(SELF);
// Look ahead
int sim_move = generate_behavior(MIRROR);
pushSimHist(sim_move);

w = mind reading window
T = mind reading threshold
simHist = simulated opponent's history
(Array of w elements)
oppHist = actual opponent's history
(Array of w elements)
```

Simply stated, this behaviour examines whether within a fixed backward looking window the opponent does what the agent would have done in its place, by simulating a copy of itself against its actual self. The function `generate_behaviour` does what the agent normally does (for example it is a TFT behaviour) and takes an argument that shows whether the inputs correspond to what the agent sees (SELF) or to what its opponent sees (MIRROR). The situation is slightly more complicated, because an agent cannot know what *exactly* its opponent sees but can only judge based on what he *thinks* its opponent sees.

This simple mind reading facility has led retaliating, even suspicious STFT-like behaviours, to converge to mutual cooperation with other agents of the same kind. Furthermore, for low values of noise (up to 10%) it has practically solved the mis-perception/mis-interpretation problem, where normally

cooperative agents find it hard to cooperate when their perception is distorted or when output moves of all agents may be distorted, respectively (distortion means switching of a move to *DEFECT*).

The condition `(opp_like_me >= T)` may be translated as `(opponent is like me)`. It is straightforward to think that what looks as a reasonable behaviour

```
if (like me) cooperate
else play as usually (reason)
```

may not be and is usually not the case. The cooperation problem may thus be defined at a meta level as:

```
if (like me) do something
else do something else
```

For example, it is not uncommon to meet people who are more cooperative with unsimilar ones (as a precaution) than with similar ones (which might be a selfish reaction). Another often encountered feature is of lower attentiveness in case of interaction with similar ones, so that signs of defection or cheating may go unnoticed for a long period.

4 CONCLUSION

Testing for mind reading to test for intelligence may have a number of assorted consequences. First, the opponent simulation part of the previous section may be incomplete. Indeed, infants have been found to somehow develop "like-me" behaviour, so normally we should endow our agents with a limited simulation possibility that is enriched in the process. Because, normally, an intelligent artifact continuously develops its own behaviour further, it makes sense to allow the mind-reading behaviour to try to copy regular behaviour, but always staying behind in complexity and performance. Second, partly due to the previous reason, mind reading is expected to be less developed than its regular counterpart. Thus externally perceived canonical intelligence may differ substantially from externally perceived social intelligence. Finally, differentially organized mind reading apparatus may be externally perceived as defective, thus giving room to the design of artificially defective agents.

In the same vein and inspired from the original Turing test, mind reading considerations bring us to consider human-machine interaction where the human will assign intelligence to the machine, but, say, childish intelligence or schizophrenic intelligence (for an old account of the latter idea see [17]). In sum, a mind reading test for intelligence allows a broadening of the scope of intelligence tests, so as to also encompass developmentally immature, defective or perceptibly distorted intelligence.

REFERENCES

[1] A. Turing, "Computing machinery and intelligence", Mind 39:433-460, 1950.
[2] A.P. Saygin, I. Cicekli, V. Akman, "Turing Test: 50 years later", Minds and Machines, 10:463-518, 2000.
[3] A. Klin, W. Jones, R. Schultz, F. Volkmar, "The enactive mind, or from actions to cognition: Lessons from autism", in U. Frith and E. Hill, Eds., "Autism: Mind and brain", Oxford University Press, 2003.
[4] S. Hurley, N. Chater, Eds., *Perspectives on imitation: From neuroscience to social science*, MIT Press, 2005.
[5] K.R. Stueber, *Rediscovering empathy: Agency, fold psychology and the human sciences*, MIT/Bradford Books, 2006.
[6] A. Turing, "Intelligent machinery", National Physical Laboratory Report, 1948.
[7] A.N. Meltzoff, "Imitation and other minds: The 'Like Me' Hypothesis", in S. Hurley, N. Chater, Eds., *Perspectives on imitation: From neuroscience to social science*, MIT Press, 2005.
[8] R. Axelrod, *The evolution of cooperation*. Basic Books, 1984.
[9] M. A. Nowak and K. Sigmund, "Tit-for-tat in heterogeneous populations", *Nature*, Vol. 355, pp. 250-53, 1992.
[10] D. Fogel, "Evolving behaviors in the iterated prisoner's dilemma", *Evolutionary Computation*, Vol. 1, pp. 77-97, 1987.
[11] M. W. Feldman and E. A. C. Thomas, "Behavior-dependent contexts for repeated plays of the prisoner's dilemma II: Dynamical aspects of the evolution of cooperation", *Journal of Theoretical Biology*, Vol 128, pp. 297-315, 1987.
[12] E. A. Stanley, D. Ashlock and L. Tesfatsion, "Iterated prisoner's dilemma with choice and refusal of partners", in *Artificial Life III*, Addison-Wesley, 1994.
[13] D. Kraines and V. Kraines, "Evolution of learning among Pavlov strategies in a competitive environment with noise", *Journal of Conflict Resolution*, Vol. 39, Issue 3, pp. 439-466, 1995.
[14] P. Molander, "The optimal level of generosity in a selfish, uncertain environment", *Journal of Conflict Resolution*, Vol. 31, Issue 4, pp. 692-724, 1987.
[15] E. Tzafestas, "Toward adaptive cooperative behavior", Proceedings of the Simulation of Adaptive Behavior Conference, Paris, September 2000.
[16] E. Tzafestas, "Attraction and cooperation in space", Proceedings 2007 Conference on Evolutionary Computation.
[17] K.M. Colby, F.D. Hilf, S. Weber, H. Kraemer, "Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes", *Artificial Intelligence*, 3:199-221, 1972..

LE PETIT CHALLENGE

Graham Wallis

(MBDA UK Ltd, Six Hills Way Stevenage SG1 2DA)

1. Abstract

The challenge takes its inspiration from the Grand Challenges, mounted successively by US DoD and UK MoD. These invoked enormous organisational resources and became very focussed on hardware, and the reliability of hardware. This smaller version is devised to focus on the AI aspects.
The primary output should be the discovery of **information management architecture**(s) which allow machines to adequately perceive and objectify their environment. This will support

Designers of Urban surveillance systems
Designers of robots for building search, street search, tunnel and cave search.
Designers of Homeland Defence systems
Designers of Autonomous Ground Vehicles and micro Air vehicles

An architecture which meets the challenge will attract attention from all users of moving imagers, and release a wave of exploitation without specific investment

2. The Challenge

There are a range of video games on the market in which the game player is “in a labyrinth”, or driving a vehicle on a racecourse or in a town. (An example might be “GoldenEye”)

The challenge is simply to replace the human game player with a second computer, programmed to move around the game purposefully. For example, to find an exit from a labyrinth, or to complete a driving route without colliding with perimeters or (static or moving) objects. The computer is to access the video stream from the host pc, and synthesise appropriate “guidance” commands to take the place of the joystick. The challenge might also include the gathering of defined objects.

Surmounting the challenge will require the development of some degree of situational awareness – in the case of the labyrinth, the ability to “know” about the walls, etc., and in the driving games the ability to interpret the moving scenery, predict the motion of the other vehicles etc

The labyrinth games have a less demanding real time constraint, as nothing happens unless the player makes a move. The driving games create a need for fast scene interpretation.

The solution requires an implementation of the OODA loop (Observe, Orient, Decide, Act); though there is a suspicion that in a human being there is an assumed orientation which precedes the observations.

3. Competition Plan

1. Seek out and select a competition organizer with a track record in related fields; example organizations based in the UK are Hertfordshire Business Incubation centre (organizing the European Galileo Masters on behalf of ESA), European Venture Contest , BBC Dragon’s Den…The PASCAL Challenge in Image Processing has recently been organized from UK universities.

2. Find software house/video gaming houses prepared to provide the technical elements – in this respect the MoD’s recent Serious Games initiatives with vendors would be explored

3. Assess several 1st person Labyrinth and driving Video Games, and select from them some examples which are relatively simple. Negotiate with the vendor of the host video game to

ensure that their Rights are not infringed, and that they will collaborate in return for the potential publicity

4. Arrange for the software house to install them on a host PC, such that the video frame store is readable by the "player" PC, and provide a method for the "player" PC to steer the game. It may be necessary to create a dedicated PCI Interface card. A fall back would be simply to arrange a web cam to view the game screen.

5. Before inviting the Rest of the World to participate, debug the approach. Objectives are to ensure that the solutions cannot be successful if tailored to the specifics of the game, that the performance can be measured progressively, etc

6. Devise a method of assessing the results and choosing a winner. Determine the timescale and decide whether the competition will held at an “event”, or in a virtual forum

7. Establish a sponsorship/prize fund by contacts with Industry stakeholders in the technology, the Venture Capital community, other investment sources such as NESTA, etc. Terms for ownership of the Intellectual Property arising from the competition and any prize have to be determined and made a condition of entry

8. Once satisfied that the competition framework is robust, the Challenge is opened to all the other interested communities, via internet, academic networks, industry networks.

Proceedings of AISB '11: Towards a Comprehensive Intelligence Test
Dimitar Kazakov and George Tsoulas (eds.)
ISBN 978-1-908187-08-6

Published by the Society for the Study of Artificial Intelligence and the Simulation of Behaviour
Printed by the University of York, York, UK