TOWARDS A UNIVERSAL PSYCHOMETRICS: Evaluating machines, animals and humans

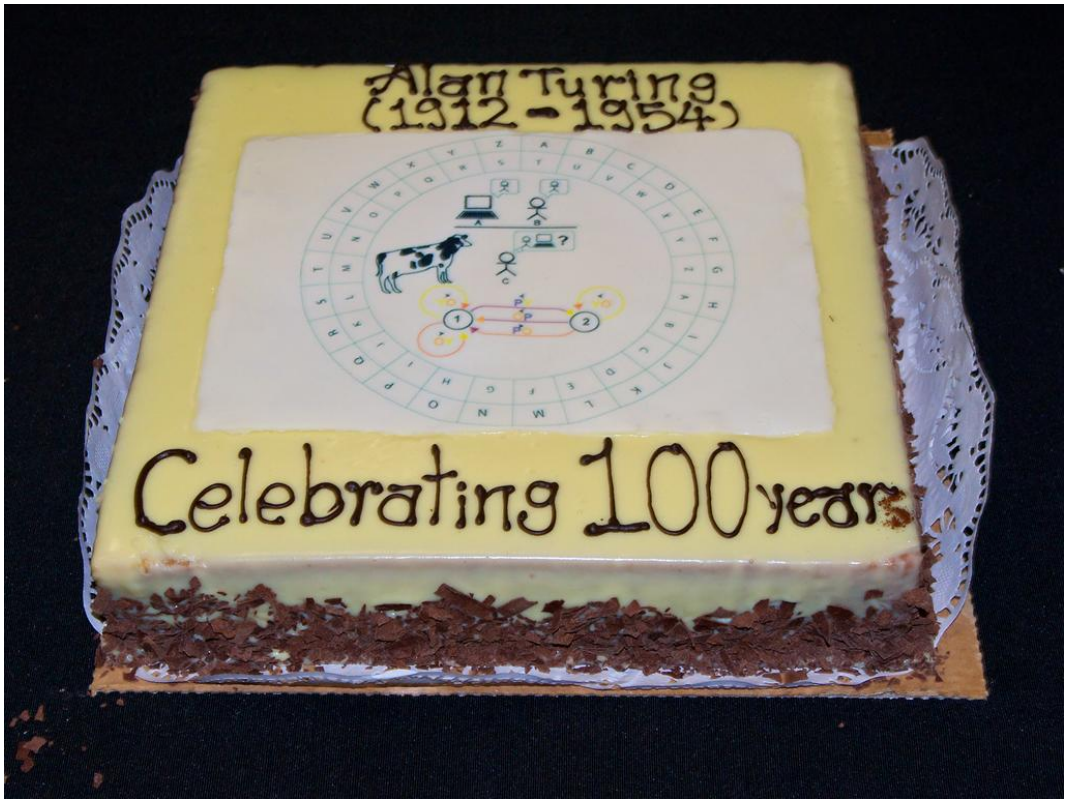
José Hernández Orallo
Dep. de Sistemes Informàtics i Computació,
Universitat Politècnica de València
jorallo@dsic.upv.es

ATENEO de la Escuela de Ingeniería y Arquitectura,
Universidad de Zaragoza, 7-Nov-2012

CELEBRATING THE ALAN TURING YEAR

STILL CELEBRATING THE ALAN TURING YEAR

- The sweetest celebration of them all!



- Cake design by David Dowe at Monash University (supported by Joy Reynolds Graphic Design, http://www.joyreynoldsdesign.com/)

OUTLINE

1. Evaluating (Turing) machines
2. Turing's Imitation Game (a.k.a. Turing Test)
3. Ca(p)tching up
4. The anthropocentric approach: psychometrics
5. Let's get chimpocentric! The animal kingdom
6. Machine evaluation beyond the Turing Test
7. Anytime universal tests
8. Universal psychometrics
9. Exploring the machine kingdom

EVALUATING (TURING) MACHINES

*Artificial Intelligence (AI) deals with the **construction** of intelligent machines.*

- *Why* is ***measuring*** important for AI?
  - Measuring and evaluation: at the roots of science and engineering.
  - Disciplines progress when they have *objective* evaluation tools to:
    - Measure the elements and objects of study.
    - Assess the prototypes and artefacts which are being built.
    - Assess the discipline as a whole.
  - Distinctions, equivalences, degrees, scales and taxonomies can be determined theoretically (on occasions), but *measuring* is the means when objects become complex, multi-faceted or physical.

EVALUATING (TURING) MACHINES

- How do other disciplines measure?
  - E.g., aeronautics: deals with the construction of flying devices.
    - Measures: mass, speed, altitude, time, consumption, load, wingspan, etc.
      - "Flying" can be defined in terms of the above measures.
    - Different specialised devices can be developed by setting different requirements over these measures.
      - Supersonic aircrafts,
      - Ultra-light aircrafts,
      - Cargo aircrafts,
      - ...

EVALUATING (TURING) MACHINES

- *What* do we want to measure in AI?
  - Algorithms? = Turing machines (Church-Turing thesis)
  - Universal Turing Machines?
  - Resource-bounded machines?
  - Physical interactive machines?
  - In actual or virtual worlds?
  - With sensors and actuators (i.e., robots)?
- The spectrum is becoming richer and richer…

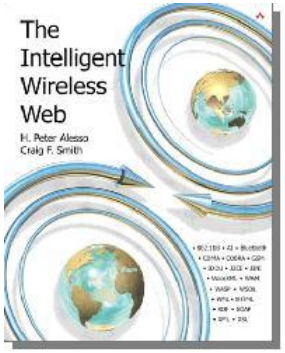
EVALUATING (TURING) MACHINES

Autonomous robots

Pets, animats and other artificial companions

Domotic systems

Agents, avatars, chatbots



Web-bots, Smartbots, Security bots…

Intelligent assistants

EVALUATING (TURING) MACHINES

- What *instruments* do we have today to evaluate all of them?

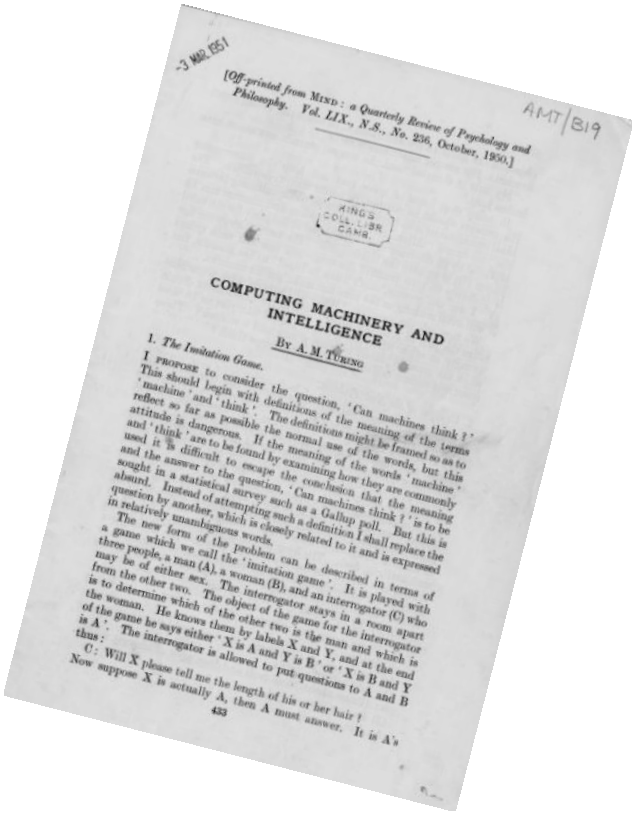
Almost nothing really general and effective !

- Why?
  - Non-biological (artificial) intelligent systems still have very limited capabilities.
    - It doesn't (or didn't) seem an imperative problem.
  - Anthropocentric formulation of AI:
    - "[AI is] the science of making machines do things that would require intelligence if done by humans." --Marvin Minsky (1968).
  - Some contests (e.g., Loebner test) have shown that non-intelligent machines can ace at these tests.

Main reason: this is a very complex problem.

TURING'S IMITATION GAME (A.K.A. TURING TEST)

- *Turing 1950: "Computing Machinery and Intelligence"*
  - "I propose to consider the question, "Can machines think?""
  - "[...] I believe to be too meaningless to deserve discussion."
    - Because he is convinced that machines *will* think.
    - Also, do *collectives* think?



-3 MAR 1951

AMT/B19

[Off-printed from MIND: a Quarterly Review of Psychology and Philosophy. Vol. LIX., N.S., No. 236, October, 1950.]

KINGS COLLEGE CAMB.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's

433

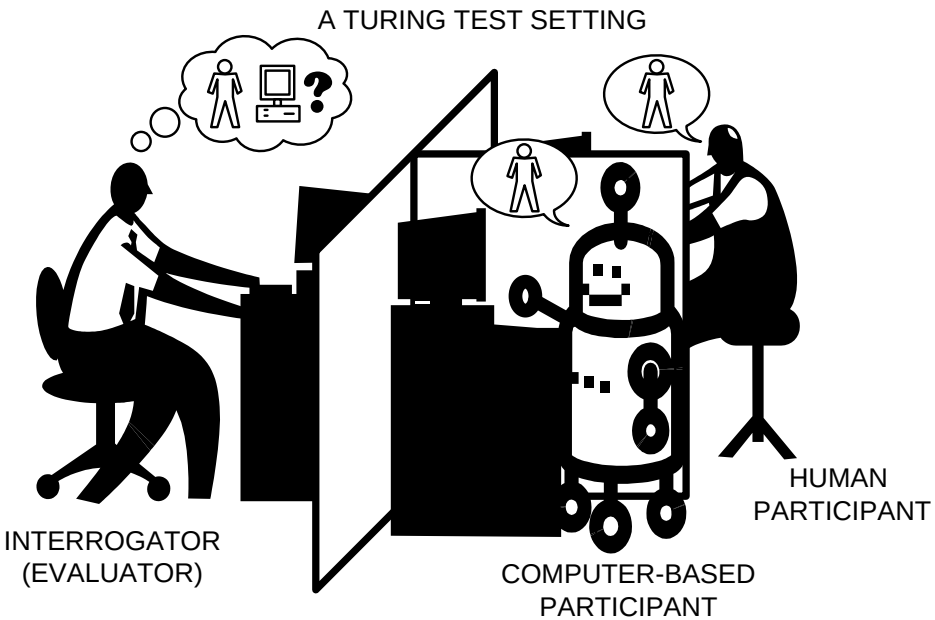
TURING'S IMITATION GAME (A.K.A. TURING TEST)

- His answer to the objections to intelligent machines is the best part of the paper, and a must-read.

  - (1) The Theological Objection -> God, souls, ...
  - (2) The "Heads in the Sand" Objection -> Dangerous machines...
  - (3) The Mathematical Objection -> Gödel, incomputability, ...
  - (4) The Argument from Consciousness -> Feelings, ...
  - (5) Arguments from Various Disabilities -> Humour, Love, Mistakes, ...
  - (6) Lady Lovelace's Objection -> Machines are programmed, they do not learn...
  - (7) Argument from Continuity in the Nervous System -> Machines are discrete...
  - (8) The Argument from Informality of Behaviour -> Humans are unpredictable...
  - (9) The Argument from Extrasensory Perception -> Mysteries in the brain...

TURING'S IMITATION GAME (A.K.A. TURING TEST)

- He *also* introduced an "imitation game":
  - A machine (A), a woman (B), and a human interrogator.
- Commonly understood as:
  - A machine (A), a human (B), and a human interrogator.
  - Known as the ***Turing Test***:



TURING'S IMITATION GAME (A.K.A. TURING TEST)

- A test?
  - It has many problems as an intelligence test:
    - It is a test of humanity, **relative** to human characteristics.
    - Neither gradual nor factorial.
    - Needs human intervention (it can't be automated).
    - It takes too much time.
    - It is anthropocentric.
    - Not a sufficient condition.
    - Not a necessary condition (a lot of human knowledge is needed).
- Turing is not to be blamed!
  - Not actually conceived by Turing to be a practical test to measure intelligence up to and beyond human intelligence.
  - A great impact in the philosophy and understanding of machine intelligence, but a *negative* impact on its measurement.

TURING'S IMITATION GAME (A.K.A. TURING TEST)

- Loebner Prize:

One sample transcript:

J: where do you work?
P: At the university. I mostly clean the Standish Building.
J: What university?
P: University of Eastern Ontario. I'm a cleaner. I empty trash. Vacuum. Dust. You know.
J: and why do they want to fire you?
P: It's just because my boss, Mr. Johnson, is mad at me.
J: why is he mad at you?
P: I didn't do nothing wrong. But he thinks I did. It's all because of that radio they think I stole from the lab.
J: that's too bad. are you in a union?
P: What would I tell the union? They won't say they're firing me because I stole the radio. They'll just make up some excuse

J is the human judge and *P* is the program

TURING'S IMITATION GAME (A.K.A. TURING TEST)

- Standard Turing Test (Loebner Prize):
  - It is becoming more difficult (more time is needed) to tell humans and machines apart.
    - Chatbots are becoming better conversation pals, but they are not becoming more intelligent (not even more human).

- Enhanced Turing Tests:
  - Total Turing Tests, Visual Turing Tests, …: including sensory information, robotic interfaces, virtual worlds, etc.
    - What about blind people (or other disabilities)?

CA(P)TCHING UP

- Artificial Intelligence: gradually catching up (and then outperforming) humans' performance for more and more tasks:
  - Calculation: 1940s-1950s
  - Cryptography: 1930s-1950s
  - Simple games (noughts and crosses, connect four, ...): 1960s
  - More complex games (draughts, bridge): 1970s-1980s
  - Data analysis, statistical inference, 1990s
  - Chess (Deep Blue vs Kasparov): 1997
  - IQ tests: 2003
  - Speech recognition: 2000s (in idealistic conditions)
  - Printed (non-distorted) character recognition: 2000s
  - TV Quiz (Watson in Jeopardy!): 2011
  - Driving a car: 2010s
  - Texas hold 'em poker: 2010s
  - Translation: 2010s (technical documents)
  - ...

No system does (or learns to do) all these things!

CA(P)TCHING UP

- Specific domain competitions:
  - Herbrand Award (automated deduction)
  - The reinforcement learning competition
  - Robocup (robot football/soccer)
  - International Aerial Robotics Competition (pilotless aircraft)
  - DARPA Grand Challenge (driverless cars)
  - NIST Face Recognition Grand Challenge
  - The planning competition
  - General game playing AAAI competition
  - BotPrize (videogame player) contest
  - Hutter Prize for Lossless Compression of Human Knowledge
  - ..

CA(P)TCHING UP

- Zadeh's Machine Intelligence Quotient (MIQ) (Zadeh 1976):
  - "MIQ –as a metric of machine intelligence– is product-**specific** and does not involve the same dimensions as human IQ. Furthermore, MIQ is **relative**, Thus, the MIQ of, say, a camera made in 1990 would be a measure of its intelligence **relative** to cameras made during the same period, and would be much lower than the MIQ of cameras made today" (Zadeh 2010, *emphasis mine*).

International Journal of Man-Machine Studies

Volume 8, Issue 3, May 1976, Pages 249–291

ELSEVIER

A fuzzy-algorithmic approach to the definition of complex or imprecise concepts †

L.A. Zadeh‡

Computer Science Division, Department of Electrical Engineering and Computer Sciences, and the Electronics Research Laboratory, University of California, Berkeley, California 94720, U.S.A.

http://dx.doi.org/10.1016/S0020-7373(76)80001-6, How to Cite or Link Using DOI

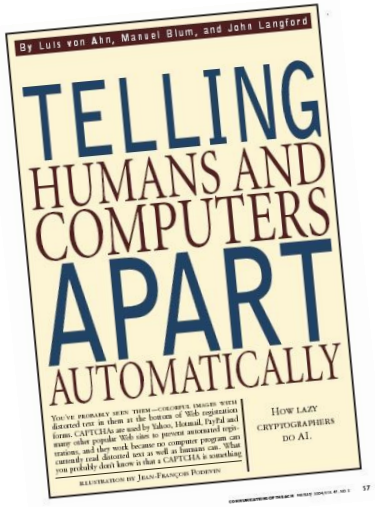
Permissions & Reprints

It may be argued, rather persuasively, that most of the concepts encountered in various domains of human knowledge are, in reality, much too complex to admit of simple or precise definition. This is true, for example, of the concepts of recession and utility in economics; schizophrenia and arthritis in medicine; stability and adaptivity in system theory; sparseness and stiffness in numerical analysis; grammaticality and meaning in linguistics; performance measurement and correctness in computer science; truth and causality in philosophy; intelligence and creativity in psychology; and obscenity and insanity in law.

The approach described in this paper provides a framework for the definition of such concepts through the use of fuzzy algorithms which have the structure of a branching questionnaire. The starting point is a relational representation of the *definiendum* as a composite question whose constituent questions are either attributional or classificational in nature. The constituent questions as well as the answers to them are

CA(P)TCHING UP

- CAPTCHAs, Completely Automated Public Turing test to tell Computers and Humans Apart (von Ahn, Blum and Langford 2002):
  - Tasks which are not in the previous lists are used to tell humans and computers apart automatically!



  - Quick and practical, omnipresent nowadays.
  - Relative to the previous list.
  - CAPTCHAs will become obsolete *in the future* (as the list evolves).
  - They are not conceived to evaluate intelligence, but to tell humans and machines apart with the current state of AI technology.

CA(P)TCHING UP

- Is there a correlation between the tasks AI is able to solve and intelligence?
  - Many of the most challenging problems for AI:
    - speech recognition, distorted character recognition, musical abilities, navigation, spatial orientation, summarisation, ...
  - can be performed almost equally well by humans of all levels of intelligence.
    - Many of them can even be performed by many animals.

- Are then AI artefacts today more *intelligent* than those of, e.g., 20 years ago?
  - In terms of general intelligence, there is no *way* to say yes.

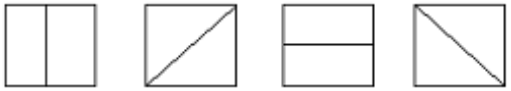
THE ANTHROPOCENTRIC APPROACH: PSYCHOMETRICS

- Goal: evaluate the intellectual abilities of human beings
  - Developed by Binet, Spearman and many others at the end of the XIXth century and first half of the XXth century.
    - Culture-fair: no "idiots savants".
  - A joint index is determined, known as IQ (Intelligence Quotient).
    - **Relative** to a population: initially normalised against the age, then normalised (μ=100, σ=15) against the adult average.
  - Tests are factorised.
    - g factor (general intelligence),
    - verbal comprehension,
    - spatial abilities,
    - memory,
    - inductive abilities,
    - calculation and deductive abilities

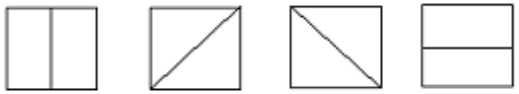
THE ANTHROPOCENTRIC APPROACH: PSYCHOMETRICS

- IQ tests are easy to administer, fast and accurate.
  - Used by companies and governments, essential in education and pedagogy.
- IQ tests are generally culture-fair through the use of abstract exercises:
  - Examples:

Consider the sequence



Which one of the following will be next in the sequence?



A B C D

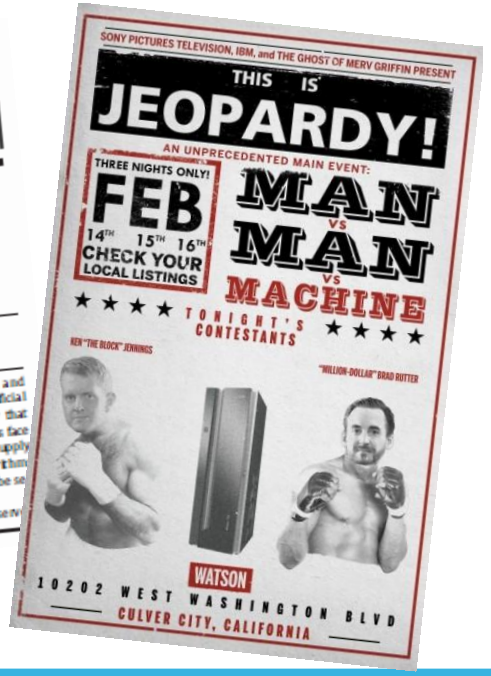
Complete the matrix

2	4	8
3	6	12
4	8	?

(except for the verbal comprehension abilities)

THE ANTHROPOCENTRIC APPROACH: PSYCHOMETRICS

- Let's use them for machines!
  - This has been suggested several times in the past
- Detterman, editor of the *Intelligence Journal*, made this suggestion serious and explicit: "A challenge to Watson (2011)"
  - As a response to specific domain tests and landmarks (such as Watson).



THE ANTHROPOCENTRIC APPROACH: PSYCHOMETRICS

- Hold on!
  - In 2003, Sanghi & Dowe implemented a small program (in Perl) which could score relatively well on many IQ tests.
    - A 3rd year student project
    - Less than 1000 lines of code

Test	I.Q. Score	Human Average
A.C.E. I.Q. Test	108	100
Eysenck Test 1	107.5	90-110
Eysenck Test 2	107.5	90-110
Eysenck Test 3	101	90-110
Eysenck Test 4	103.25	90-110
Eysenck Test 5	107.5	90-110
Eysenck Test 6	95	90-110
Eysenck Test 7	112.5	90-110
Eysenck Test 8	110	90-110
I.Q. Test Labs	59	80-120
Testedich.de:I.Q. Test	84	100
I.Q. Test from Norway	60	100
Average	96.27	92-108

*This made the point unequivocally: this program is **not intelligent***

THE ANTHROPOCENTRIC APPROACH: PSYCHOMETRICS

- Rejoinder:
  - "IQ tests are not for machines" (Dowe & Hernandez-Orallo 2012)
    - IQ tests take many things for granted:
      - They are anthropocentric.
        - More than that, they are specialised to the average human.
        - Tests are broader when evaluating small children, people with disabilities, etc.?
    - We can devise different IQ test batteries such that AI systems (e.g., Sanghi and Dowe's program) fail:
      - This would end up as a psychometric CAPTCHA.

Contents lists available at SciVerse ScienceDirect

Intelligence

journal homepage:

IQ tests are not for machines, yet

David L Dowe [a,*], José Hernández-Orallo [b,*]

[a] Computer Science and Software Engineering, Clayton School of Information Technology, Monash University, Clayton, Vic. 3800, Australia

[b] DSIC, Universitat Politècnica de València, Valencia, Spain

ARTICLE INFO

Article history:
Received 20 September 2011
Received in revised form 3 November 2011
Accepted 22 December 2011
Available online xxxx

Keywords:
Machine intelligence evaluation
IQ tests
Artificial intelligence
Universal tests
Psychometrics
Task difficulty
CAPTCHA

ABSTRACT

Complex, but specific tasks—such as chess or *Jeopardy!*—are popularly seen as milestones for artificial intelligence (AI). However, they are not appropriate for evaluating the intelligence of machines or measuring the progress in AI. Aware of this delusion, Detterman has recently raised a challenge prompting AI researchers to evaluate their artefacts against IQ tests. We agree that the philosophy behind (human) IQ tests is a much better approach to machine intelligence evaluation than these specific tasks, and also more practical and informative than the Turing test. However, we have first to recall some work on machine intelligence measurement which has shown that some IQ tests can be passed by relatively simple programs. This suggests that the challenge may not be so demanding and may just work at a sophisticated CAPTCHA, since some types of tests might be easier than others for the current state of AI. Second, we show that an alternative, formal derivation of intelligence tests for machines is possible, grounded in (algorithmic) information theory. In these tests, we have a proper mathematical definition of what is being measured. Third, we re-visit some research done in the past fifteen years for effectively measuring machine intelligence—since some assumptions about the subjects and their distribution no longer hold.

1. Introduction: The challenge

In February 2011, Douglas K. Detterman announced a challenge (originally to IBM's program Watson (Ferrucci et al., 2010), the recent winner of the *Jeopardy!* TV quiz show at the time) for the whole field of artificial intelligence (AI). AI artefacts should be better measured by classical IQ tests. The challenge goes as follows (Detterman, 2011): "I, the editorial board of *Intelligence*, and members of the International Society for Intelligence Research will develop a unique battery of intelligence tests that would be administered to that computer and would result in an actual IQ score".

Computers are (still) so stupid today, that it seems clear that an average result at IQ tests is far beyond current computer technology. "It is doubtful that anyone will take up this challenge in the near future", Detterman said (Detterman, 2011). But the challenge had already been taken up, in the past.

In 2003, a computer program performed quite well on standard human IQ tests (Sanghi & Dowe, 2003). This was an elementary program, far smaller than Watson or the successful chess-playing Deep Blue (Campbell, Hoane, & Hsu, 2002). The program had only about 960 lines of code in the programming language Perl (accompanied by a list of 25,143 words), but it even surpassed the average score (of 100) on some tests (Sanghi & Dowe, 2003, Table 1).

The computer program underlying this work was based on the realisation that most IQ test questions that the authors had seen until then tended to be of one of a small number of types or formats. Formats such as "insert missing letter/number in middle or at end" and "insert suffix/prefix to complete two or more words" were included in the program. Other formats such as "complete matrix of numbers/characters", "use directions, comparisons and/or pictures", "find the odd

* Corresponding author at: DSIC, Universitat Politècnica de València, Camí de Vera s/n, 46022 València, Spain. Tel.: +34 96 3877007x75465; fax: +34 96 3877359.

E-mail addresses: david.dowe@infotech.monash.edu.au (D.L. Dowe), jorallo@dsic.upv.es (J. Hernández-Orallo).

0160-2896/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2011.12.001

LET'S GET CHIMPOCENTRIC! THE ANIMAL KINGDOM

- Chimpanzees:

FROM: Herrmann, E., Call, J., Hernández-Lloreda, M.V., Hare, B., Tomasello, M. "Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis", Science, 7 September 2007, Vol. 317. no. 5843, pp. 1360 - 1366, DOI: 10.1126/science.1146282.

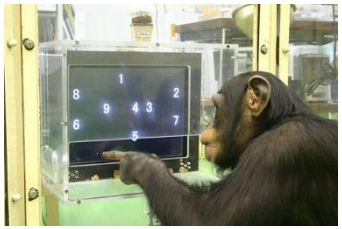
LET'S GET CHIMPOCENTRIC! THE ANIMAL KINGDOM

- Human children:

FROM: Herrmann, E., Call, J., Hernández-Lloreda, M.V., Hare, B., Tomasello, M. "Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis", Science, 7 September 2007, Vol. 317. no. 5843, pp. 1360 - 1366, DOI: 10.1126/science.1146282.

LET'S GET CHIMPOCENTRIC! THE ANIMAL KINGDOM

- Animal evaluation and comparative psychology
  - How are tests conducted?
    - Use of rewards
    - Relevance of interfaces
    - Animals and compared (abilities are "**relative to…**")
  - Is it isolated from psychometrics?
    - Partly it was, but it is becoming closer and closer, especially when comparing apes and human children
    - Many abilities which were considered exclusively human have been found in many animals.



Images from BBC One documentary: "Super-smart animal": http://www.bbc.co.uk/programmes/b01by613

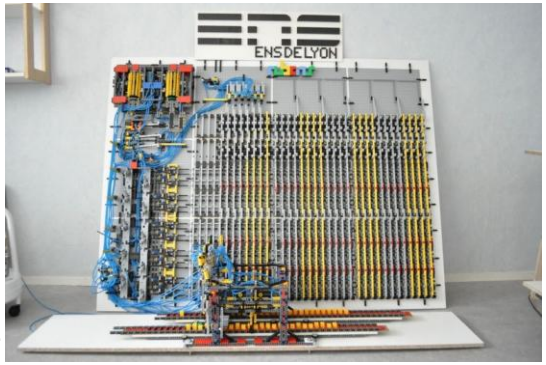
LET'S GET CHIMPOCENTRIC! THE ANIMAL KINGDOM

- Is it applicable to machines?
  - The selection of tasks and abilities is not systematic.
    - Some tasks would be too easy for machines (e.g., memory).
    - Others would be difficult (e.g., orientation, recognition, interaction).
  - But many ideas (and the overall perspective) are useful:
    - Abilities as concepts.
    - Tests as instruments.
    - Rewards and interfaces.
    - Testing social abilities (co-operation and competition) is common.
    - No prejudices.
    - Non-anthropocentric:
      - exploring the animal kingdom.
      - humans as a special case.

MACHINE EVALUATION BEYOND THE TURING TEST

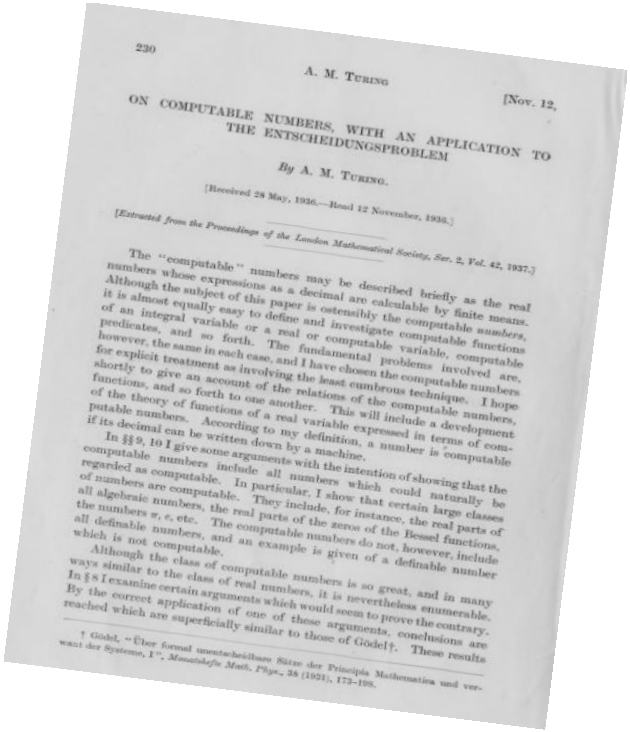
- A different approach to machine evaluation started in the late 1990s
  - Back to Turing (not 1950, but 1936!)
  - (Universal) Turing Machines.

10111100100 <-



<- 01000100100

Lego Turing machine. Rubens project



230 A. M. TURING [Nov. 12,

ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM

By A. M. TURING.

[Received 28 May, 1936.—Read 12 November, 1936.]

[Extracted from the Proceedings of the London Mathematical Society, Ser. 2, Vol. 42, 1937.]

The "computable" numbers may be described briefly as the real numbers whose expressions as a decimal are calculable by finite means. Although the subject of this paper is ostensibly the computable numbers, it is almost equally easy to define and investigate computable functions of an integral variable or a real or computable variable, computable predicates, and so forth. The fundamental problems involved are, however, the same in each case, and I have chosen the computable numbers for explicit treatment as involving the least cumbrous technique. I hope shortly to give an account of the relations of the computable numbers, functions, and so forth to one another. This will include a development of the theory of functions of a real variable expressed in terms of computable numbers. According to my definition, a number is computable if its decimal can be written down by a machine.

In §§ 9, 10 I give some arguments with the intention of showing that the computable numbers include all numbers which could naturally be regarded as computable. In particular, I show that certain large classes of numbers are computable. They include, for instance, the real parts of all algebraic numbers, the real parts of the zeros of the Bessel functions, the numbers π, e, etc. The computable numbers do not, however, include all definable numbers, and an example is given of a definable number which is not computable.

Although the class of computable numbers is so great, and in many ways similar to the class of real numbers, it is nevertheless enumerable. In § 8 I examine certain arguments which would seem to prove the contrary. By the correct application of one of these arguments, conclusions are reached which are superficially similar to those of Gödel†. These results

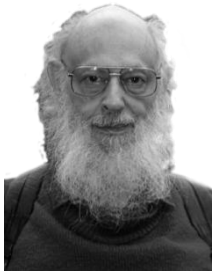
† Gödel, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandt der Systeme, I", *Monatshefte Math. Phys.*, 38 (1931), 173–198.

Based on (algorithmic) information theory, compression, inductive inference, probability, …

MACHINE EVALUATION BEYOND THE TURING TEST

A. M. Turing (1936), (Universal) Turing machines, Church-Turing thesis

C. E. Shannon (1948), information theory, connection between probability and information



R. J. Solomonoff (1964): algorithmic information theory and algorithmic probability.

A. N. Kolmogorov (1965), probability axioms, independent development of algorithmic information theory

G. J. Chaitin (1969, 1966) works on algorithmic information theory, mathematics, life complexity.

CS Wallace and DM Boulton (1968), MML principle, information theory and two-part compression for (statistical) inference.

MACHINE EVALUATION BEYOND THE TURING TEST

- Kolmogorov complexity, $K_U(s)$: shortest program for machine U which describes/outputs an object s (e.g., a binary string).
- Algorithmic probability (universal distribution), $p_U(s)$: the probability of objects as outputs of a UTM U fed by 0/1 from a fair coin.
- Both are related (under prefix-free or monotone TMs):

$$p_U(s) = 2^{-K_U(s)}$$

- Invariance theorem: the value of $K(s)$ (and hence $p(s)$) for two different reference UTMs U_1 *and* U_2 only differs by (at most) a constant (which is independent of s).
  - *Hence*, these measures are usually said to be '**absolute**' (up to a constant).
- $K(s)$ is incomputable, but approximations exist (Levin's Kt).

MACHINE EVALUATION BEYOND THE TURING TEST

- Many variants for different views of complexity (and difficulty): logical depth, sophistication, average case computational complexity, …
- Formalisation of Occam's razor: shorter is better!
- Compression and inductive inference (and *learning*): two sides of the same coin (Solomonoff, MML, …).
- Its direct relation to intelligence measurement occasionally suggested:
  - "measuring machine power-intelligence as the scope of the class of inferable functions" (Blum and Blum, 1975).
  - "develop formal definitions of intelligence and measures of its various components [using algorithmic information theory]" (Chaitin 1982)
  - "what kind of information-processing is intelligence?" (Chandrasekaran 1990).

MACHINE EVALUATION BEYOND THE TURING TEST

- Compression and intelligence
  - Compression-enhanced Turing Tests (Dowe & Hajek 1997-1998).
    - A Turing Test which includes compression problems.
    - By ensuring that the subject needs to compress information, we can make the Turing Test more sufficient as a test of intelligence and discard objections such as Searle's Chinese room.

A Computational Extension to the Turing Test

David L. Dowe and Alan R. Hájek

Department of Computer Science, Monash University, Clayton, Vic. 3168, Australia
HSS, California Institute of Technology, Pasadena, California 91125, U.S.A.
e-mail: {dld@cs.monash.edu.au, ahajek@hss.caltech.edu}

August 17, 1997

Abstract

The purely behavioural nature of the Turing Test leaves many with the view that passing it is not sufficient for 'intelligence' or 'understanding'. We propose here an additional necessary computational requirement on intelligence that is non-behavioural in nature and which we contend is necessary for a commonsense notion of 'inductive learning' and, relatedly, of 'intelligence'. Said roughly, our proposal is that a key to these concepts is the notion of compression of data. Where the agent under assessment is able to communicate, e.g. by a tele-type machine, our criterion is that, in addition to requiring the agent's being able to pass Turing's original (behavioral) Turing Test, we also require that the agent have a somewhat compressed representation of the test domain. Our reason for adding this requirement is that, as we shall argue from both Bayesian and information-theoretic grounds, inductive learning and compression are tantamount to the same thing. We can only compress data when we learn a pattern or structure, and it seems quite reasonable to require that an 'intelligent' agent can inductively learn (and record the result learnt from the compression). We illustrate these ideas and our extension of the Turing Test via Searle's Chinese room example and the problem of other minds.

We also ask the following question: Given two programs H_1 and H_2 respectively of lengths l_1 and l_2, $l_1 < l_2$, if H_1 and H_2 perform equally well (to date) on a Turing Test, which, if either, should be preferred for the future?

We also set a challenge. If humans can presume intelligence in their ability to set the Turing test, then we issue the additional challenge to researchers to get machines to *administer* the Turing Test.

Keywords: Turing Test, Philosophy of AI, compression, Bayesian and Statistical Learning Methods, Machine Learning, Cognitive Modelling.

1

MACHINE EVALUATION BEYOND THE TURING TEST

- Intelligence *definition* and *test* (C-test) based on algorithmic information theory (Hernandez-Orallo 1998-2000).
  - Series are generated from a TM with a general alphabet and some properties (projectibility, stability, ...).

$k = 9$	: a, d, g, j, ...	Answer : m
$k = 12$	: a, a, z, c, y, e, x, ...	Answer : g
$k = 14$	: c, a, b, d, b, c, c, e, c, d, ...	Answer : d

Journal of Logic, Language, and Information 9: 447–465, 2000.
© 2000 Kluwer Academic Publishers. Printed in the Netherlands.

447

Beyond the Turing Test

JOSE HERNANDEZ-ORALLO
Departamento de Sistemas Informáticos y Computación, Universitat Politècnica de València, Camí de Vera, s/n, E-46022 València, Spain
E-mail: jorallo@dsic.upv.es

(Received 1 June 1999; in final form 15 April 2000)

Abstract. The main factor of intelligence is defined as the ability to comprehend, formalising this ability with the help of new constructs based on descriptional complexity. The result is a computer test, or C-test, which is explicitly defined in computational terms. Due to its absolute and non-anthropomorphic character, it is equally applicable to both humans and non-humans. Moreover, it correlates with classical psychometric tests, thus establishing the first firm connection between information theoretical notions and traditional IQ tests. The Turing Test is compared with the C-test and the combination of both is proposed. In consequence, the idea of using the Turing Test as a practical test of intelligence should be questioned. In contrast, the idea of using the Turing Test as a battery of different cognitive abilities, a much more useful approach to be substituted by computational and factorial tests, or other intriguing questions that present themselves beyond the Turing Test.

Key words: AI's anthropomorphism, comprehension, descriptional complexity, inductive inference, measurement of intelligence, psychometrics, Turing test

1. Introduction

Turing realised that only machines with very special self-modifying programs could eventually pass the test he had just devised (Turing, 1950). The ability to learn new functions, for which the machine had not explicitly been programmed has since then been recognised as a necessary condition for a machine to be considered intelligent. Accordingly, the relationships between AI, IQ tests, inductive inference, learning, and descriptional complexity soon began to be discovered and in many cases, formalised.

Many early works in AI dealt with the relation between IQ tests and AI (Simon and Kotovsky, 1963; Evans, 1968). The association between AI and inductive inference was taken on by Solomonoff (1957), leading him to the independent discovery of descriptional complexity (also known nowadays as algorithmic information or Kolmogorov complexity) and the formulation of inductive inference under this complexity (Solomonoff, 1964, 1978). Since intelligence tests occasionally require the extrapolation of an effective sequence, the connections between inductive inference and IQ tests was frequently alluded to (Gold, 1967; Blum and Blum, 1975). In particular, the idea of "measuring machine power-intelligence as the scope of the class of inferable functions" is suggested for the first time (Blum and Blum,

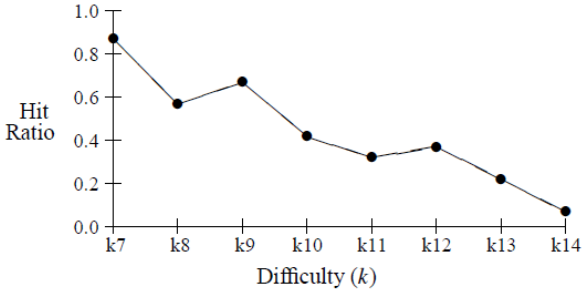
- Intelligence is the result of a test:

$$I_U(\sigma) \triangleq \sum_{e \in \mathbb{E}} \rho_U(e) \cdot w_U(e,\varepsilon) \cdot H_e^\sigma \approx \frac{1}{n'} \sum_{k=K_{min}\ldots K_{max}} k^\varepsilon \cdot \sum_{e_{j,k} \text{ with } Kt_U(e_j)=k, j=1\ldots n} H_{e_{j,k}}^\sigma$$

where U is the reference machine, σ is the subject, $\mathbb{E}$ is the set of all possible exercises (stable projectible series), ε is the weight depending on the difficulty, H_e^σ is whether the subject σ makes the exercise e correctly, and n' is a normalisation factor which depends on n, ε, K_{max} and K_{min}.

MACHINE EVALUATION BEYOND THE TURING TEST

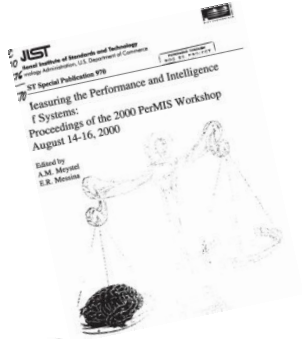
- Very much like IQ tests, but formal and well-grounded :
  - exercises are not chosen arbitrarily.
  - the right solution (projection of the sequence) is 'unquestionable'.
  - Item difficulty derived in an 'absolute' way.
- *Human performance correlated with the absolute difficulty (k) of each exercise and IQ tests for the same subjects:*



- This is IQ-test re-engineering!
  - However, some relatively simple programs can ace on them (e.g., Sanghi and Dowe 2003).
  - They are static (series): no planning/"action" required.

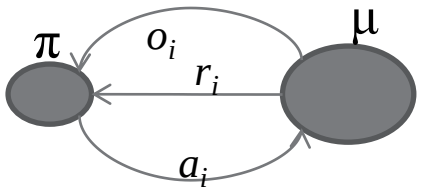
MACHINE EVALUATION BEYOND THE TURING TEST

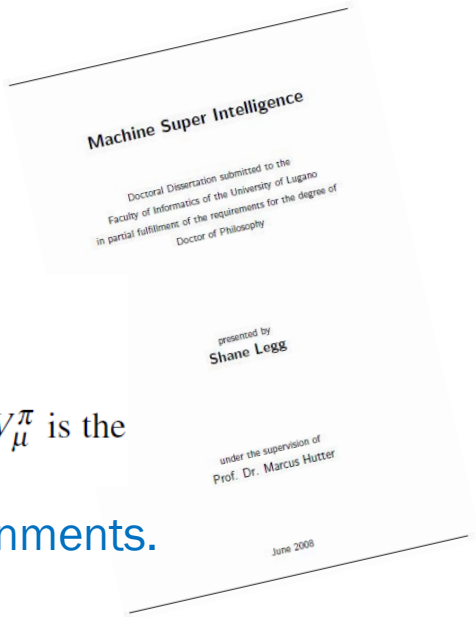
- First workshop on Performance of Machine Intelligence Systems, at the US National Institute of Standards and Technology.
  - *(Hernández-Orallo 2000b) "On the computational measurement of intelligence factors"*
    - *looking for a sufficient set of abilities*
    - *factorisation: deduction, knowledge acquisition*
    - *"rewards and penalties could be used instead", as in reinforcement learning.*
  - *(Zadeh 2000) "The search for metrics of intelligence – a critical view" argued that "a realistic metrization of intelligence is not possible within the conceptual structure of existing methods of definitions and measurement. We cannot expect a concept as complex as intelligence to be definable in traditional terms."*



MACHINE EVALUATION BEYOND THE TURING TEST

- *"Universal Intelligence"* (Legg and Hutter 2007): an interactive extension of C-tests from sequences to environments…




Machine Super Intelligence

Doctoral Dissertation submitted to the Faculty of Informatics of the University of Lugano in partial fulfillment of the requirements for the degree of Doctor of Philosophy

presented by
Shane Legg

under the supervision of
Prof. Dr. Marcus Hutter

June 2008

- Intelligence as performance over many environments.

$$\Upsilon_U(\pi) \triangleq \sum_{\mu \in \mathbb{E}} p_U(\mu) \cdot V_\mu^\pi = \sum_{\mu \in \mathbb{E}} 2^{-K_U(\mu)} \cdot E\left(\sum_{i=1}^{\infty} r_i^{\mu,\pi}\right)$$

where U is the reference machine, π is the agent, $\mathbb{E}$ is the set of all environments and V_μ^π is the expected sum of rewards of π in μ.

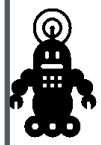
  - The mass of the probability measure goes to a few environments.
  - The probability distribution is not computable.
  - Most environments are not really discriminative.
  - There are two infinite sums (number of environments and interactions).
  - Time/speed is not considered for the environment or for the agent.

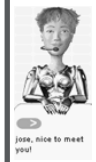
ANYTIME UNIVERSAL TESTS

- Machine intelligence evaluation at the dawn of the XXIst century…
  - *Fascinating* but… *discouraging* state:
    - We still have no effective intelligence test for machines.
    - Scattered efforts:
      - on different areas, with different philosophies, tools, foundations, terminologies, …
      - on different kinds of subjects to be evaluated.
    - Not even recognised as an imperative problem.
    - Certainly not a mainstream area of research.

ANYTIME UNIVERSAL TESTS

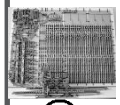
- A snapshot of the fragmentation…



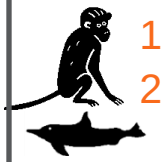


- Turing test:
  1. Held in a human natural language.
  2. The examinees 'know' it is a test.
  3. Interactive.
  4. Adaptive.
  5. Relative to humans.

- IQ tests:
  1. Human-specific tests.
  2. The examinees know it is a test.
  3. Generally non-interactive.
  4. Generally non-adaptive (pre-designed set of exercises)
  5. Relative to a population



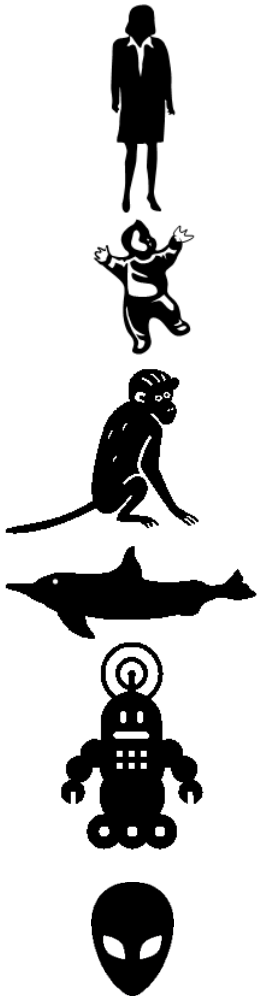
- Tests and definitions based on AIT
  1. Interaction highly simplified.
  2. The examinees do not know it is a test. Rewards may be used.
  3. Sequential or interactive.
  4. Non-adaptive.
  5. Formal foundations.



- Animal (and children) intelligence evaluation:
  1. Perception and action abilities assumed.
  2. The examinees do not know it is a test. Rewards are used.
  3. Interactive.
  4. Generally non-adaptive.
  5. Comparative (relative to other species)

Other task-specific tests: robotics, games, machine learning.

ANYTIME UNIVERSAL TESTS



- Can we construct a test for all of them?
  - Without knowledge about the examinee,
  - Derived from computational principles,
  - Non-biased (species, culture, language, etc.)
  - No human intervention,
  - Producing a score,
  - Meaningful,
  - Practical, and
  - Anytime.

ANYTIME UNIVERSAL TESTS

- Anytime universal test (Hernandez-Orallo & Dowe 2010):
  - The class of environments is carefully selected to be discriminative.
  - Environments are randomly sampled from that class.
    - Starts with very simple environments.
    - Complexity of the environments adapts to the subject's performance.
  - The speed of interaction adapts to the subject's performance.
  - Includes time.
  - It can be stopped anytime.

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

ELSEVIER

Measuring universal intelligence: Towards an anytime intelligence test

José Hernández-Orallo [a,*], David L. Dowe [b]

[a] Departamento de Sistemas Informáticos y Computación, Universitat Politècnica de València, Camí de Vera s/n, E-46022, València, Spain
[b] Computer Science & Software Engineering, Clayton School of I.T., Monash University, Clayton, Victoria, 3800, Australia

ARTICLE INFO

Article history:
Received 16 December 2009
Received in revised form 24 September 2010
Accepted 24 September 2010
Available online xxxx

Keywords:
Measurement of intelligence
Artificial intelligence
Psychometrics
Algorithmic information theory
Kolmogorov complexity
Algorithmic probability
Turing test
Universal intelligence
Computerized adaptive testing
Compression
Inductive inference
Prediction
Minimum Message Length (MML)
Reinforcement learning

ABSTRACT

In this paper, we develop the idea of a universal anytime intelligence test. The meaning of the terms "universal" and "anytime" is manifold here: the test should be able to measure the intelligence of any biological or artificial system that exists at this time or in the future. It should also be able to evaluate both inept and brilliant systems (any intelligence level) as well as very slow to very fast systems (any time scale). Also, the test may be interrupted at any time, producing an approximation to the intelligence score, in such a way that the more time is left for the test, the better the assessment of machine intelligence based proposal is based on previous works on the measurement of machine intelligence based on Kolmogorov complexity and universal distributions, which were developed in the late 1990s (C-tests and compression-enhanced Turing tests). It is also based on the more recent idea of measuring intelligence through dynamic/interactive tests and highlights their limitations distribution of environments. We discuss some of these tests and highlight their limitations since we want to construct a test that develop early "compression tests" and the more recent introduce many new ideas that develop early "compression tests" and the more recent definition of "universal intelligence" in order to design new "universal intelligence tests", where a feasible implementation has been a design requirement. One of these tests is the "anytime intelligence test", which adapts to the examinee's level of intelligence in order to obtain an intelligence score within a limited time.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Artificial intelligence has arguably given preference to the goal of developing intelligent systems over the goal of evaluating them. The evaluation of artificial intelligence has almost always been addressed in the context of what artificial intelligence should be, or even in the context of whether it was possible. In fact, we are uncertain of how much the field of artificial intelligence has progressed in precise measurable terms since its inception. However, the eventual achievement of true and general artificial intelligence does not entail, per se, that intelligence can be evaluated. Nor does a definition of intelligence entail that intelligence can be evaluated with it. From an engineering point of view, the most desirable sequence of events would be to have a definition, then a feasible measurement procedure, and from these, to build intelligent systems that could be ultimately certified and evaluated with the measurement procedure.

The evaluation and measurement of intelligence is related to the notion of test, which is the daily basis of psychometrics; however, it is also the approach taken by the so-called Turing test [1], its variants, and extensions. In artificial intelligence, the major concern has been more about when and how our artefacts will be able to pass any of these tests, and not as

* Corresponding author.
E-mail addresses: jorallo@dsic.upv.es (J. Hernández-Orallo), David.Dowe@infotech.monash.edu.au (D.L. Dowe).

ANYTIME UNIVERSAL TESTS

- The test is an adaptive algorithm:

Definition 18 *(Anytime universal intelligence test taking time into account).* We define $\Upsilon^{\nu}(\pi, U, H, \Theta)$ as the result of the following algorithm, which can be stopped anytime:

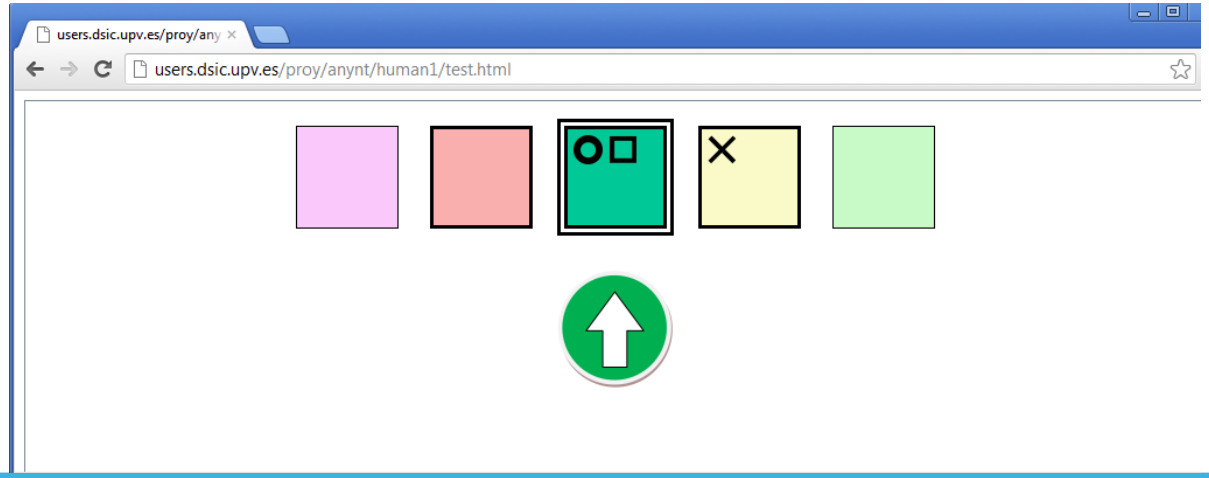
```
ALGORITHM: Anytime Universal Intelligence Test
INPUTS: π (an agent), U (a universal machine), H (a complexity function),
        Θ (test time, not as a parameter if the test is stopped anytime)
OUTPUTS: a real number (approximation of the agent's intelligence)
BEGIN
 Υ ← 0                                  (initial intelligence)
 τ ← 1 microsecond                      (or any other small time value)
 ξ ← 1                                  (initial complexity)
 S_used ← ∅                             (set of used environments, initially empty)
 WHILE (TotalElapsedTime < Θ) DO
  REPEAT
    μ ← Choose(U, ξ, H, S_used)         (get a balanced, reward-sensitive environment with ξ − 1 ⩽ H ⩽ ξ not already in S_used)
   IF (NOT FOUND) THEN                  (all of them have been used already)
    ξ ← ξ + 1                           (we increment complexity artificially)
   ELSE
    BREAK REPEAT                        (we can exit the loop and go on)
   END IF
  END REPEAT
  Reward ← V_μ^π ‖ τ                    (average reward until time-out τ stops)
  Υ ← Υ + Reward                        (adds the reward)
  ξ ← ξ + ξ · Reward/2                  (updates the level according to reward)
  τ ← τ + τ/2                           (increases time)
  S_used ← S_used ∪ {μ}                 (updates set of used environments)
 END WHILE
 Υ ← Υ/|S_used|                         (averages accumulated rewards)
 RETURN Υ
END ALGORITHM
```

ANYTIME UNIVERSAL TESTS

- The anYnt project (2009-2011):

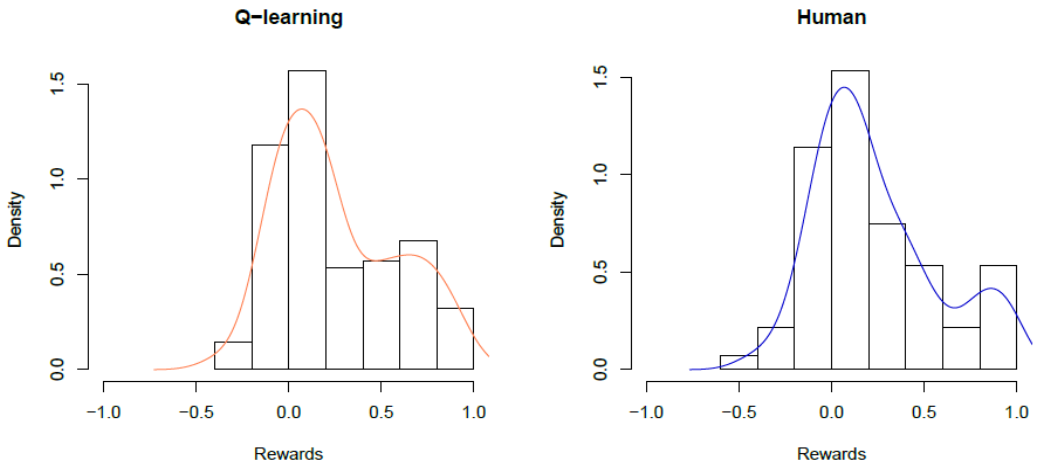
  http://users.dsic.upv.es/proy/anynt/

  - Goal: evaluate the feasibility of a universal test.
    - What do environments look like? An environment class Λ was devised.
    - The complexity/difficulty function Kt^{max} was chosen.
    - An interface for humans was designed.



ANYTIME UNIVERSAL TESTS

- Experiments (2010-2011):
  - The test is applied to humans and an AI algorithm (Q-learning):



  - Impressions:
    - The test is useful to compare and scale systems of the same type.
    - The results do not reflect the actual differences between humans and Q-learning.

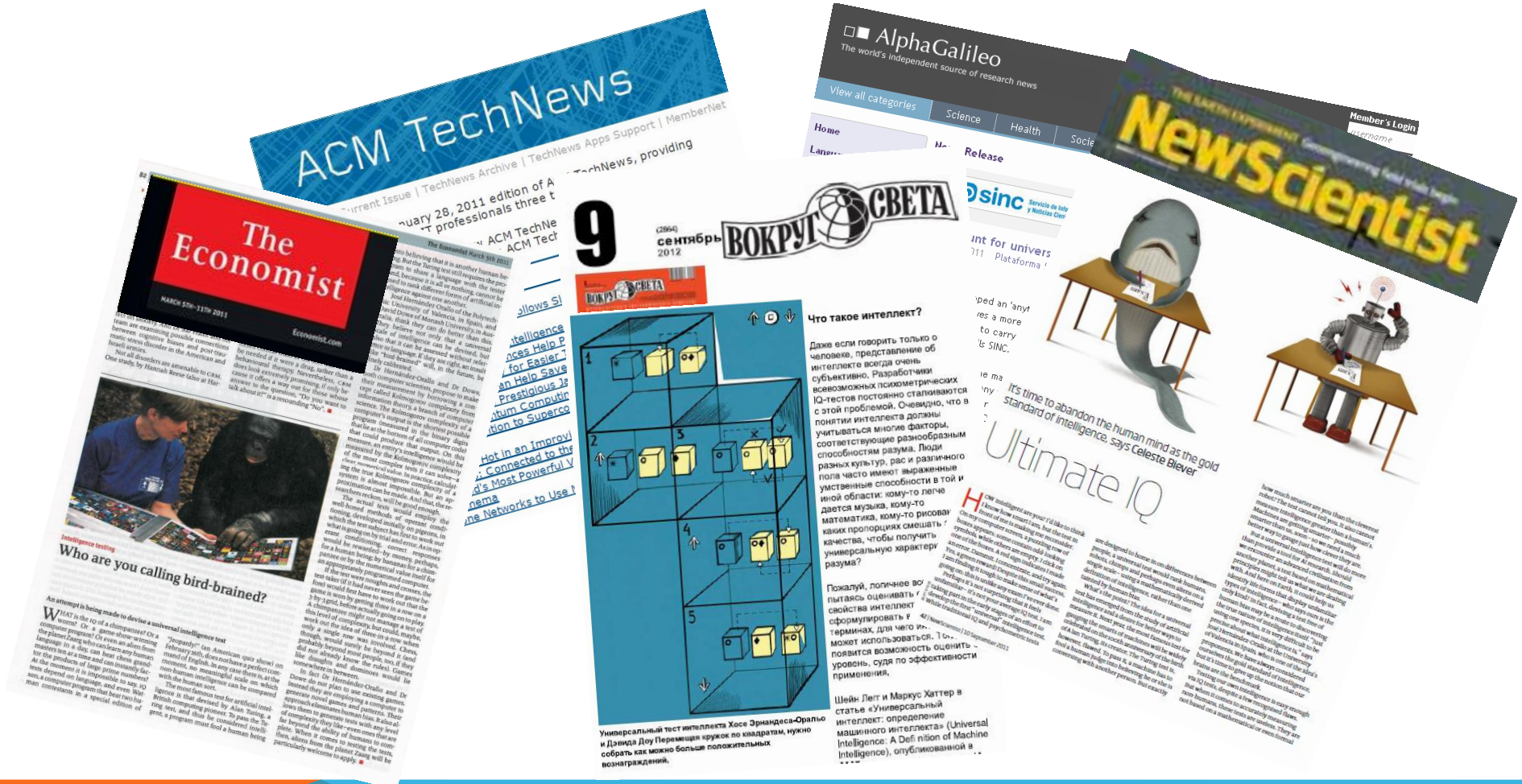
ANYTIME UNIVERSAL TESTS

An intelligence test, based on theoretical principles, applied to humans and machines.

- How should this be interpreted?
  - It was a prototype: many simplifications made.
  - It is not adaptive (not anytime)
  - Absence of noise: specially beneficial for AI agents.
  - Patterns have low complexity.
  - The environment class may be richer.
  - More factors may be needed.
  - No incremental knowledge acquisition.
  - No social behaviour (environments weren't multi-agent).
- Are universal tests impossible?
  - All the above issues should be explored before dismissing this idea.

ANYTIME UNIVERSAL TESTS

- anYnt project media coverage! (despite the limited results)



ANYTIME UNIVERSAL TESTS

- *Something went very wrong here…*

UNIVERSAL PSYCHOMETRICS

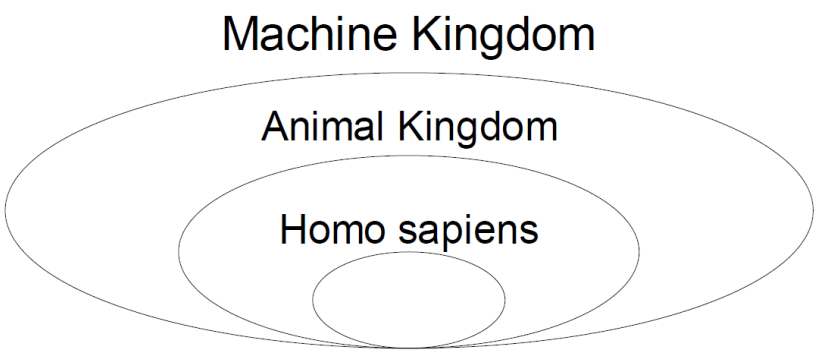
- Evaluation is always harder the less we know about the subject.
  - The less we take for granted about the subjects the more difficult it is to construct a test for them.
    - Human intelligence evaluation (psychometrics) works because it is highly specialised for humans.
    - Animal testing works (relatively well) because tests are designed in a very specific way to each species.

Who would try to tackle a more general problem (evaluating *any* system) instead of the actual problem (evaluating machines)?

UNIVERSAL PSYCHOMETRICS

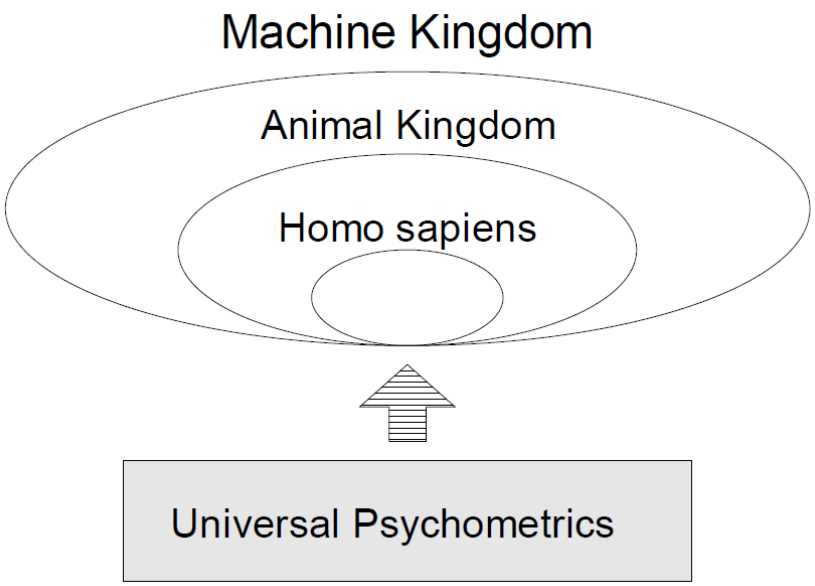
- The *actual* problem is the *general* problem:
  - What about 'animats'? And hybrids? And collectives?

Machine kingdom: any kind of individual or collective, either artificial, biological or hybrid.



UNIVERSAL PSYCHOMETRICS

Universal Psychometrics is the analysis and development of measurement techniques and tools for the evaluation of cognitive abilities of subjects in the machine kingdom.



Universal Psychometrics:
Measuring Cognitive Abilities in the Machine Kingdom

José Hernández-Orallo
DSIC, Universitat Politècnica de València, València, Spain.
jorallo@dsic.upv.es
David L. Dowe
Computer Science and Software Engineering, Clayton School of Information Technology
Monash University, Vic. 3800, Australia.
david.dowe@infotech.monash.edu.au
M.Victoria Hernández-Lloreda
Departamento de Metodología de las Ciencias del Comportamiento
Universidad Complutense de Madrid, Spain.
vallorea@psi.ucm.es

October 1, 2012

Abstract

We present and develop the notion of 'universal psychometrics' as a subject of study, and eventually a discipline, that focuses on the measurement of cognitive abilities for the machine kingdom, which comprises any (cognitive) system, individual or collective, either artificial, biological or hybrid. Universal psychometrics can be built, of course, upon the experience, techniques and methodologies from (human) psychometrics, comparative cognition and related areas. Conversely, the perspective and techniques which are being developed in the area of machine intelligence measurement using (algorithmic) information theory can be of much broader applicability and implication outside artificial intelligence. This general approach to universal psychometrics spurs the re-understanding of most (if not all) of the big issues about the measurement of cognitive abilities, and creates a new foundation for (re)defining and mathematically formalising the concept of cognitive task, evaluable subject, interface, task choice, difficulty, agent response curves, etc. We introduce the notion of a universal cognitive test and discuss whether (and when) it may be necessary for exploring the machine kingdom. On the issue of intelligence and very general abilities, we also get some results and connections with the related notions of no-free-lunch theorems and universal priors.

Keywords: cognitive system evaluation; psychometrics; cognitive abilities; artificial intelligence; algorithmic information theory

1 Introduction

Much of what we know scientifically about intelligence and cognitive abilities in general today originates from psychometrics and related disciplines, where measures and measurement procedures have been their tools and objects of study. Psychometrics has been successful in developing valid and reliable tests for human intelligence and related abilities. However, psychometrics has not given

UNIVERSAL PSYCHOMETRICS

- Elements:
  - Subjects: physically computable (resource-bounded) interactive systems.
  - Cognitive task: physically computable interactive systems with a score function.
  - Interfaces: between subjects and tasks (observations-outputs, actions-inputs), score-to-reward mappings.
  - Distributions over a task class
    - performance as average case performance on a task class.
  - Difficulty functions computationally defined from the task itself.
    - difficulty for each single task, not for the task class.
- Some of them present in psychometrics and, most especially, comparative cognition, but we must overhaul them here with the theory of computation and algorithmic information theory.

UNIVERSAL PSYCHOMETRICS

- Intelligence in psychometrics and comparative psychology is usually seen as:
  - "what intelligence tests measure" (Boring 1923).
- In universal psychometrics:
  - Cognitive abilities can be seen as classes of tasks, perfectly defined in computational terms.
    - The relation between abilities can be explored experimentally, but also theoretically.
  - Measures are absolute and not relativised wrt. a population.
    - Except for social abilities (competition and co-operation).
  - Tests can be universal or not, depending on the application.
- Strong objections are understandable, given the 'failure' of machine intelligence evaluation in the past 60 years.

EXPLORING THE MACHINE KINGDOM

- *Explorers* needed!
  - The machine kingdom is a space of cosmic dimension!

> "A smart machine will first consider which is more worth its while: to perform the given task or, instead, to figure some way out of it. Whichever is easier. And why indeed should it behave otherwise, being truly intelligent? For true intelligence demands choice, internal freedom. And therefore we have the malingerants, fudgerators, and drudge-dodgers, not to mention the special phenomenon of simulimbecility or mimicretinism. A mimicretin is a computer that plays stupid in order, once and for all, to be left in peace. And I found out what dissimulators are: they simply pretend that they're *not* pretending to be defective. Or perhaps it's the other way around. The whole thing is very complicated."
>
> Stanislaw Lem, "The Futurological Congress (1971)"

EXPLORING THE MACHINE KINGDOM

- Intelligence measurement is still an open problem.
  - But it is arguably the most important piece for understanding what intelligence is (and, of course, to devise intelligent artefacts).
  - Already needed in some applications (CAPTCHAs, social networks, certification, etc.)
  - More and more common in the future: plethora of bots, robots, artificial agents, avatars, control systems, 'animats', hybrids, collectives, etc.
  - Crucial for the *technological singularity* once (and if) achieved.
- The exploration of the machine kingdom is dual to the exploration of the set of possible cognitive abilities/tasks.
  - As in the theory of computation: e.g., *problem* classes and *automata* classes.

EXPLORING THE MACHINE KINGDOM

- Our early motivation was the lack of proper intelligence measurements for machines.
  - This motivation is strengthened and refined:

> Artificial intelligence requires an accurate, non-anthropocentric, meaningful and computational way of evaluating its progress, by evaluating its artefacts.

> Evaluating machine intelligence must be seen as a very general problem, subsuming (and relating to) many other previous approaches to intelligence evaluation.

- Turing (1950):
  - "We can only see a short distance ahead, but we can see plenty there that needs to be done."

THANK YOU!

- Special thanks to David Dowe,
  - and the rest of members of the anYnt project:

    http://users.dsic.upv.es/proy/anynt/

  - for their joint work, ideas, material, software, experiments, patience and support:
    - M.Victoria Hernández-Lloreda,
    - Javier Insa,
    - Sergio España.

- And also to http://www.turingarchive.org for Turing's original papers, and Greg Chaitin, Douglas Hofstadter, Marcus Hutter and Shane Legg for (re-)invigorating the will for working in this area (in different ways and at different times in the past fifteen years).