MEASURING (MACHINE) INTELLIGENCE UNIVERSALLY:

An interdisciplinary challenge*

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

9th Congress of the EUS-UES
Valencia, 15-17 October 2014

* A paper closely related to this presentation, including full coverage of topics and references, can be found at: http://arxiv.org/abs/1408.6908

to ***Rafael Beneyto Torres***,
in memoriam

OUTLINE

- Understanding “Intelligent” Systems
- Task-oriented evaluation
  - Types of performance measurement
- Towards ability-based evaluation
  - What is an ability?
  - The anthropocentric approach: psychometrics
  - The information-theoretic approach
  - Universal psychometrics
- Discussion
- Conclusions

WHAT COMES TO MIND WITH “MACHINE INTELLIGENCE”?

Image from wikicommons

UNDERSTANDING "INTELLIGENT" SYSTEMS

- AI systems are pervading our life and our workplace.
  - Increasingly more tasks are performed by AI systems:
    - from sorting our emails to driving a car.
  - Many decisions are made by or using AI systems.
    - in companies, organisations and governments.
- AI systems have become very complex.
  - Integrate many different technologies and knowledge.
  - Evolve from the interaction with users and other machines.

AI systems are unpredictable

  - Unpredictable at the action (low) level!
    - No longer like (old) software systems.
    - White-box approaches are insufficient.

UNDERSTANDING "INTELLIGENT" SYSTEMS

- Systemic (black-box) approaches are necessary.
  - We analyse "properties" of these systems (designed, emergent, unexpected)
  - Experimental validation is more common.
- Can we predict their behaviour in terms of more abstract properties?
  - Performance, abilities, psychological traits, personality, …
- The type of intelligent system is very important:
  - Task-oriented AI systems
  - Ability-oriented AI systems

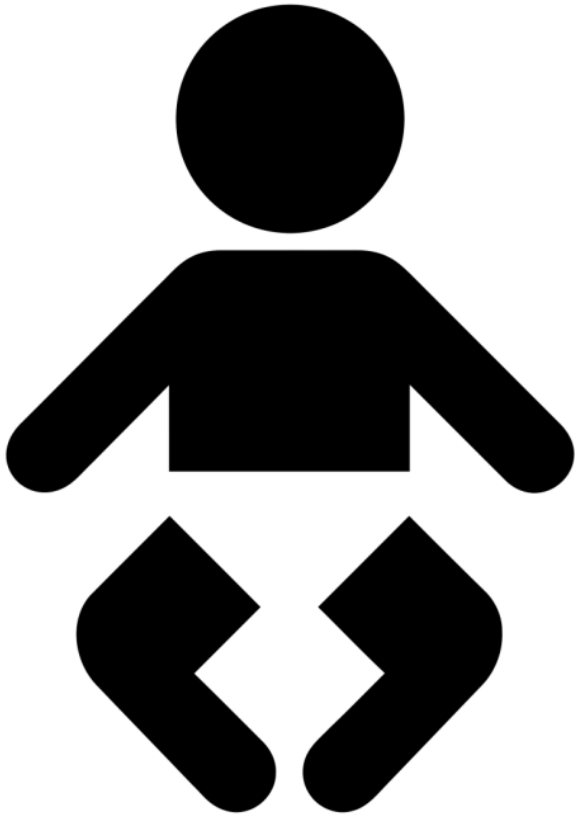
UNDERSTANDING “INTELLIGENT” SYSTEMS

Task-oriented

(clear functionality)

Image from wikicommons

UNDERSTANDING “INTELLIGENT” SYSTEMS



Warning!
Completely useless until grown up.

ability-oriented

(no functionality)

Image from wikicommons

UNDERSTANDING "INTELLIGENT" SYSTEMS

- Some AI systems are also devised to be general, and solve a variety of tasks or to assist humans and organisations.
  - Robots, bots, avatars and 'smart' devices are enhancing our capabilities as individuals, collectives and humanity.

- Many kinds of systems:
  - Pure AI systems.
    - learning systems, expert systems, …
  - Hybrid individuals:
    - enhanced human intelligence, human computing.
  - Collective intelligence:
    - humans and machines, crowd computing.

UNDERSTANDING "INTELLIGENT" SYSTEMS

- What are these systems capable of doing?
- What is their intelligence?
- How to tell whether they are meeting their "specifications"?
- Are the organisations using them being less predictable and governable?

We lack proper measurement tools to evaluate the cognitive abilities and expected behaviour of this variety of systems.

TASK-ORIENTED EVALUATION

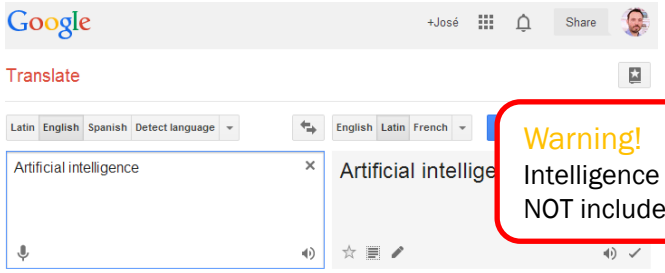
- Specific (task-oriented) AI systems

Warning!
Intelligence
NOT included.

Computer vision, speech recognition, etc.)

Warning!
Intelligence
NOT included.

Prediction and estimation



Warning!
Intelligence
NOT included.

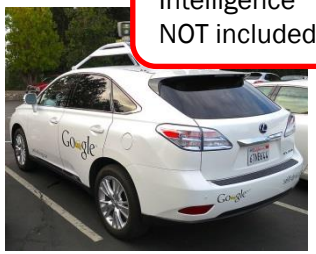
Machine translation, information retrieval, summarisation

Warning!
Intelligence
NOT included.

Robotic navigation

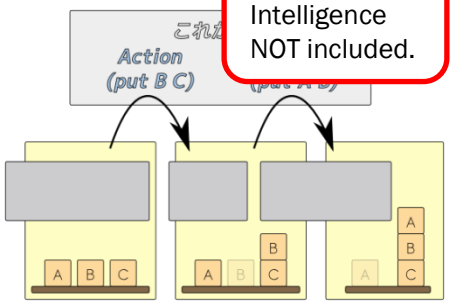
Warning!
Intelligence
NOT included.

Expert systems



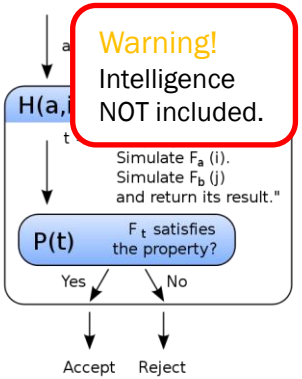
Warning!
Intelligence
NOT included.

Driverless vehicles



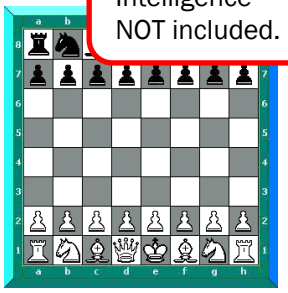
Warning!
Intelligence
NOT included.

Planning and scheduling



Warning!
Intelligence
NOT included.

Automated deduction



Warning!
Intelligence
NOT included.

Game playing

All images from wikicommons

TASK-ORIENTED EVALUATION

- *How is AI producing so many non-intelligent systems?*
  - *In disagreement with the ambitious view of AI:*

> *"[Artificial Intelligence (AI) is] the science and engineering of making intelligent machines." —John McCarthy (2007)*

  - *In agreement with a more pragmatic view of AI:*

> *"[AI is] the science of making machines do things that would require intelligence if done by [humans]." —Marvin Minsky (1968).*

    - Machines need not be intelligent!
    - They can do the "things" (tasks) **without featuring intelligence.**

TASK-ORIENTED EVALUATION

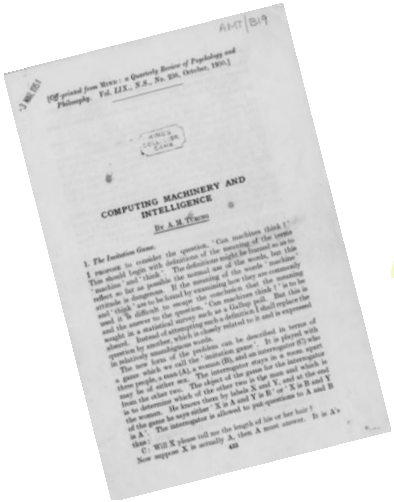
- Consider:
  - A set of problems, tasks or exercises, M.
  - For each exercise $\mu \in M$, we can get a measurement $R(\pi, \mu)$ of the performance of system π.
    - We will use $E[R(\pi, \mu)]$ when the system, the problem or the measurement is non-deterministic and/or imperfect.

- Three common types of aggregated performance metrics:
  - Worst-case performance:
    - $\Phi_{min}(\pi, M) = \min_{\mu \in M} E[R(\pi, \mu)]$
  - Best-case performance:
    - $\Phi_{max}(\pi, M) = \max_{\mu \in M} E[R(\pi, \mu)]$
  - Average-case performance:
    - $\Phi(\pi, M, p) = \sum_{\mu \in M} p(\mu) \cdot E[R(\pi, \mu)]$
      - where $p(\mu)$ is a probability distribution on M.

TYPES OF PERFORMANCE MEASUREMENT IN AI

- Human discrimination (observation, scrutiny and/or interview):
  - Assessment is made by and/or against humans. Usually informal.
  - Common in psychology, ethology and comparative psychology.
  - Not usual in AI (except for the Turing Test and variants).

- Examples: Loebner Prize, Bot Prize, …


COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

TYPES OF PERFORMANCE MEASUREMENT IN AI

- Problem benchmarks:
  - Collections or repositories (a set of problems *M* is set up).
    - Common in AI: repositories, problem libraries, corpora, etc.
    - Also usual in (comparative) psychology (e.g., cognitive tests).
    - Solutions may be known beforehand or can be inferred by humans.
    - Most systems actually embed what the researchers have learnt.

    Risk: systems overfit to the evaluation benchmarks
  - Problem generators (a class of problems is derived with a generator).
    - This actually defines *M* and *p*.
    - Better characterisation of each problem (e.g., difficulty).
  - Need to distinguish the aggregation from the evaluation procedure.

- Examples: planning competition, Arcade Learning Environment…

TYPES OF PERFORMANCE MEASUREMENT IN AI

- Peer confrontation (1-vs-1 or n-vs-n).
  - Evaluates performance in (multi-agent) games from a set of matches.
  - The result is relative to the other participants.
    - Sophisticated performance metrics (e.g., the Elo system in chess).
  - Systems can specialise to the kind of opponents that are expected in a competition. This is usual in sports.

- Examples: General Game Competition, Robocup, …

EXAMPLES OF EVALUATION SETTINGS

- Specific domain evaluation settings:
  - CADE ATP System Competition → PROBLEM BENCHMARKS
  - Termination Competition → PROBLEM BENCHMARKS
  - The reinforcement learning competition → PROBLEM BENCHMARKS
  - Program synthesis (Syntax-guided synthesis) → PROBLEM BENCHMARKS
  - Loebner Prize → HUMAN DISCRIMINATION
  - Robocup and FIRA (robot football/soccer) → PEER CONFRONTATION
  - International Aerial Robotics Competition (pilotless aircraft) → PROBLEM BENCHMARKS
  - DARPA driverless cars, Cyber Grand Challenge, Rescue Robotics → PROBLEM BENCHMARKS
  - The planning competition → PROBLEM BENCHMARKS
  - General game playing AAAI competition → PEER CONFRONTATION
  - BotPrize (videogame player) contest (2014 in Spain) → HUMAN DISCRIMINATION
  - World Computer Chess Championship → PEER CONFRONTATION
  - Computer Olympiad → PEER CONFRONTATION
  - Annual Computer Poker Competition → PEER CONFRONTATION
  - Trading agent competition → PEER CONFRONTATION
  - Robo Chat Challenge → HUMAN DISCRIMINATION
  - UCI repository, PRTools, or KEEL dataset repository. → PROBLEM BENCHMARKS
  - KDD-cup challenges and ML kaggle competitions → PROBLEM BENCHMARKS
  - Machine translation corpora: Europarl, SE times corpus, the euromatrix, Tenjinno competitions… → PROBLEM BENCHMARKS
  - NLP corpora: linguistic data consortium, … → PROBLEM BENCHMARKS
  - Warlight AI Challenge → PEER CONFRONTATION
  - The Arcade Learning Environment → PROBLEM BENCHMARKS
  - Pathfinding benchmarks (gridworld domains) → PROBLEM BENCHMARKS
  - Genetic programming benchmarks → PROBLEM BENCHMARKS
  - CAPTCHAs → HUMAN DISCRIMINATION
  - Graphics Turing Test → HUMAN DISCRIMINATION
  - FIRA HuroCup humanoid robot competitions → PROBLEM BENCHMARKS
  - …

TOWARDS ABILITY-ORIENTED EVALUATION

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

TOWARDS ABILITY-ORIENTED EVALUATION

- Tasks are classified as: (Rajani 2010, Information Technology)
  - *optimal*: it is not possible to perform better
  - *strong super-human*: performs better than all humans
  - *super-human*: performs better than most humans
  - *par-human*: performs similarly to most humans
  - *sub-human*: performs worse than most humans

- This view of "progress in artificial intelligence" is misleading.
  - All these systems are task-oriented systems.

TOWARDS ABILITY-ORIENTED EVALUATION

No AI system can do (*or can learn to do*) **all** these things!

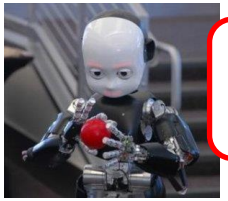
*Despite pitiful **big-switch** approaches*

But this system can:

A different perspective for AI evaluation:
"machines do tasks they have never seen and have not been prepared for beforehand."

TOWARDS ABILITY-ORIENTED EVALUATION

- How can we evaluate more general AI systems?



Warning!
Some intelligence
MAY BE included.

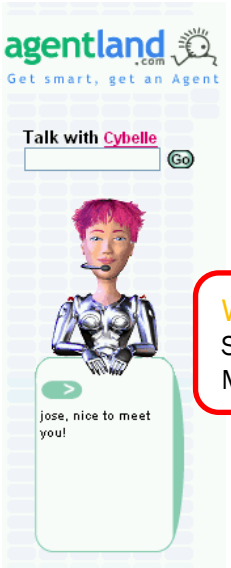
Cognitive robots

Warning!
Some intelligence
MAY BE included.

Pets, animats and other artificial companions

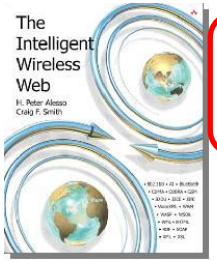
Warning!
Some intelligence
MAY BE included.

Smart buildings



Warning!
Some intelligence
MAY BE included.

Agents, avatars, chatbots



Warning!
Some intelligence
MAY BE included.

Web-bots, Smartbots, Security bots…

Warning!
Some intelligence
MAY BE included.

Intelligent assistants

WHAT IS AN ABILITY?

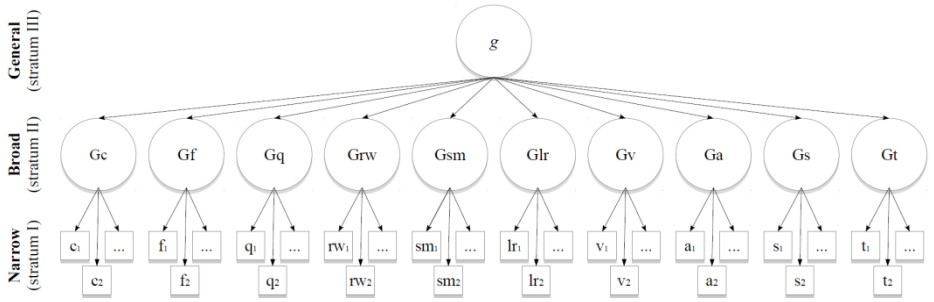
- We are talking about *cognitive* abilities:

> A cognitive ability is a property of individuals which allows them to perform well in a *range* of information-processing tasks.

  - The ability is *required*.
    - Performance is much worse ***without featuring the ability***.
    - Note that the ability is *necessary* but it does *not* have to be *sufficient*.
      - E.g., spatial abilities are necessary but not sufficient for driving a car.
  - *General*, covering a range of tasks.
  - Problem: abilities have to be conceptualised and identified.
    - Abilities are constructs while tasks are instruments.

WHAT IS AN ABILITY?

- Many arrangements of cognitive abilities have been identified.
  - For instance, the Cattell-Horn-Carroll theory:
    - Broad abilities:
      - Crystallised Intelligence (Gc), Fluid Intelligence (Gf), Quantitative Reasoning (Gq), Reading and Writing Ability (Grw), Short-Term Memory (Gsm), Long-Term Storage and Retrieval (Glr), Visual Processing (Gv), Auditory Processing (Ga), Processing Speed (Gs) and Decision/Reaction Time/Speed (Gt)

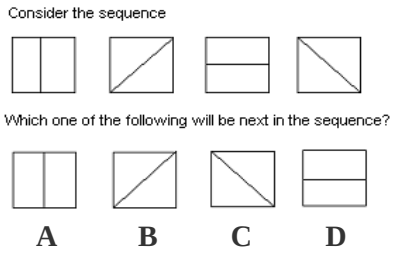


  - The broad abilities seem to correspond to subfields in AI:
    - problem solving, use of knowledge, reasoning, learning, perception, natural language processing, … (from Russell and Norvig 2009).

Figure adaptation courtesy of Fernando Martínez-Plumed

THE ANTHROPOCENTRIC APPROACH: PSYCHOMETRICS

- Goal: evaluate the intellectual abilities of human beings
  - Developed by Galton, 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 usually 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.
- IQ tests are easy to administer, fast and accurate.
  - Used by companies and governments, essential in education and pedagogy.
  - Tests are factorised.
    - g factor (general intelligence),
    - verbal comprehension,
    - spatial abilities,
    - memory,
    - inductive abilities,
    - calculation and deductive abilities

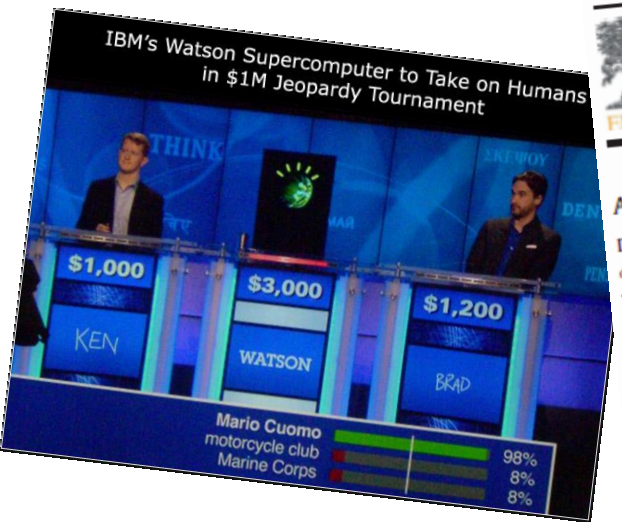


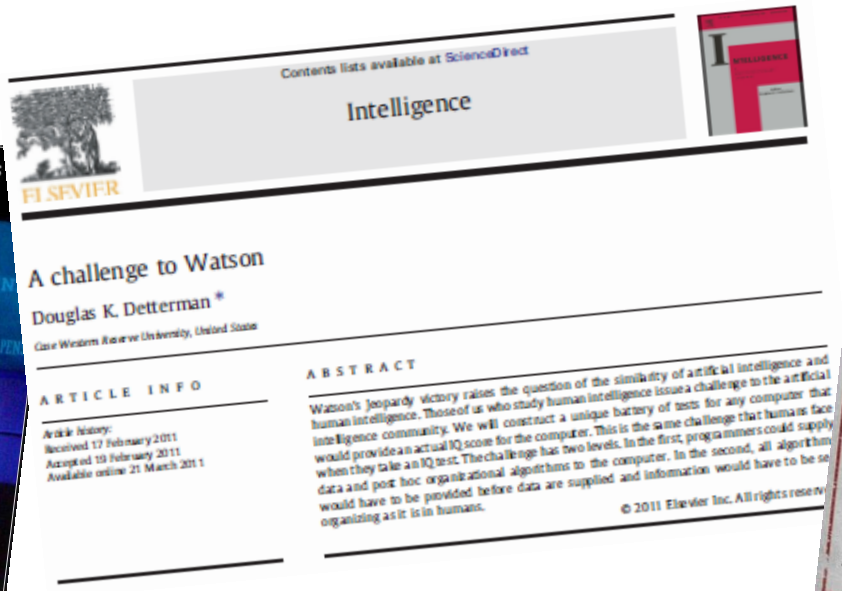
Complete the matrix

2	4	8
3	6	12
4	8	?

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).




Contents lists available at ScienceDirect

Intelligence

ELSEVIER

A challenge to Watson

Douglas K. Detterman *

Case Western Reserve University, United States

ARTICLE INFO

ABSTRACT

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
      - (a big-switch approach)

This made the point unequivocally: this program is not intelligent

Warning!
Intelligence
NOT included.

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

THE ANTHROPOCENTRIC APPROACH: PSYCHOMETRICS

- Response to Detterman:
  - "IQ tests are not for machines, yet" (Dowe & Hernandez-Orallo 2012, Intelligence Journal)
  - IQ tests take many things for granted:
    - They are anthropocentric.
      - On top of that, they are specialised to the average human.
      - Tests are broader when evaluating small children, people with disabilities, etc.?
  - Can we 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.
  - IQ tests are increasingly more used in AI
    - For a survey, Hernandez-Orallo et al. 2015, AIJ.

THE ~~ANTHROPOCENTRIC~~ CHIMPOCENTRIC APPROACH!

- Animal evaluation and comparative psychology
  - 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
  - Applicable to machines?
    - Not directly.
    - But many ideas (and the overall perspective) are useful:
      - Use of rewards and interfaces
      - Abilities as concepts and tests as instruments.
      - Testing social abilities (co-operation and competition) is common.
      - No prejudices.
      - Non-anthropocentric:
        - exploring the animal kingdom.
        - humans as a special case.

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

THE INFORMATION-THEORETIC APPROACH

- A different approach to evaluation started in the late 1990s
  - Algorithmic Information Theory (Turing, Shannon, Solmonoff, Kolmogorov, Chaitin, Wallace)
    - 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.
      - Immune to the NFL theorem (every computable distribution can be approximated by a universal distribution).
    - 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).
    - $K(s)$ is incomputable, but approximations exist (Levin's Kt).
  - Formalisation of Occam's razor: shorter is better!
  - Compression and inductive inference (and *learning*): two sides of the same coin (Solomonoff, MML, ...).

THE INFORMATION-THEORETIC APPROACH

- 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.
  - But it is still a Turing Test…

THE INFORMATION-THEORETIC APPROACH

- 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

- Intelligence is the result of a test:

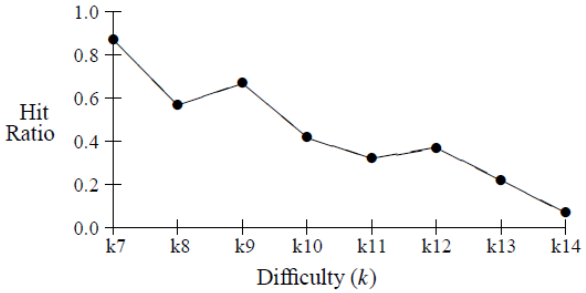
$$I_U(\sigma) \triangleq \sum_{e \in \mathbb{E}} \rho_U(e) \cdot w_U(e, \varepsilon) \cdot H_e^\sigma \approx \frac{1}{n'} \sum_{k=K_{min} \ldots K_{max}} k^\varepsilon \cdot \sum_{e_{j,k} \text{ with } Kt_U(e_j)=k, j=1 \ldots n} H_{e_{j,k}}^\sigma$$

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}.

  - Bears similarities with our aggregated measures using *M* and *p*.

THE INFORMATION-THEORETIC APPROACH

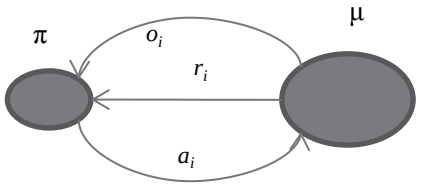
- Very much like IQ tests, but formal and well-grounded :
  - exercises are not chosen arbitrarily.
  - the right solution (projection of the sequence) is 'unquestionable'.
  - Item difficulty derived in an 'absolute' way.
    - *Human performance correlated with the absolute difficulty (k) of each exercise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



- This is IQ-test re-engineering!
  - However, some simple programs can ace on them (e.g., Sanghi and Dowe 2003).
  - They are static (series): no planning/"action" required.
  - Only covers general intelligence. Other abilities (Hernández-Orallo 2000b, NIST)

THE INFORMATION-THEORETIC APPROACH

- Intelligence as performance in a range of worlds. (Hutter 2000, Dobrev 2000, 2005)
  - Worlds: interactive environments
    - R is understood as the degree of success
  - The set of worlds *M* is described by Turing machines.
    - Bounded or weighted by Kolmogorov complexity.



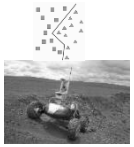
  - Intelligence is measured as an average, following the average-case evaluation:
    - $\Phi(\pi, M, p) = \sum_{\mu \in M} p(\mu) \cdot \mathrm{E}[R(\pi, \mu)]$
  - *"Universal Intelligence"* (Legg and Hutter 2007): much better formalised.
  - Both are interactive extensions of C-tests from sequences to environments...

- Problems:
  - For both approaches, the mass of the probability measure goes to a few environments.
  - *M* or the probability distribution is not computable.
  - Most environments are not really discriminative (Dobrev discusses this issue briefly).
  - (Legg and Hutter) There are two infinite sums (environments and interactions).
  - Time/speed is not considered for the environment or for the agent.

UNIVERSAL PSYCHOMETRICS

- A snapshot of the fragmentation of intelligence evaluation…

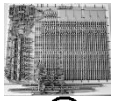
- Human-discriminative (e.g., Turing test):
  1. Held in a human natural language.
  2. The examinees 'know' it is a test.
  3. Interactive.
  4. Adaptive.
  5. Relative to humans.



- Problem benchmarks:
  1. Task-specific tests.
  2. Choice of problems not always representative.
  3. Generally non-adaptive.
  4. Risk of problem overfitting.

- Peer confrontation:
  1. Task-specific tests.
  2. Highly dependent on the opponents (relative to a population)
  3. Standard measurements difficult to obtain
  4. A good match arrangement necessary for reliability of results.

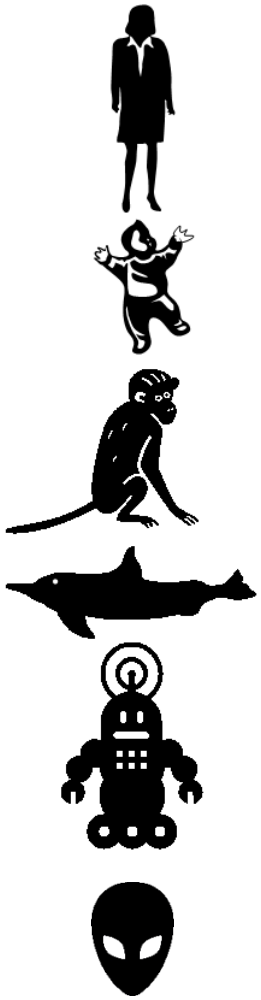
- IQ tests:
  1. Human-specific tests.
  2. The examinees know it is a test.
  3. Generally non-interactive.
  4. Generally non-adaptive (pre-designed set of exercises)
  5. Relative to a population



- 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).

UNIVERSAL PSYCHOMETRICS



- Can we construct tests 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.

- A *multidisciplinary* approach?

Need of an *interdisciplinary* systemic approach

UNIVERSAL PSYCHOMETRICS

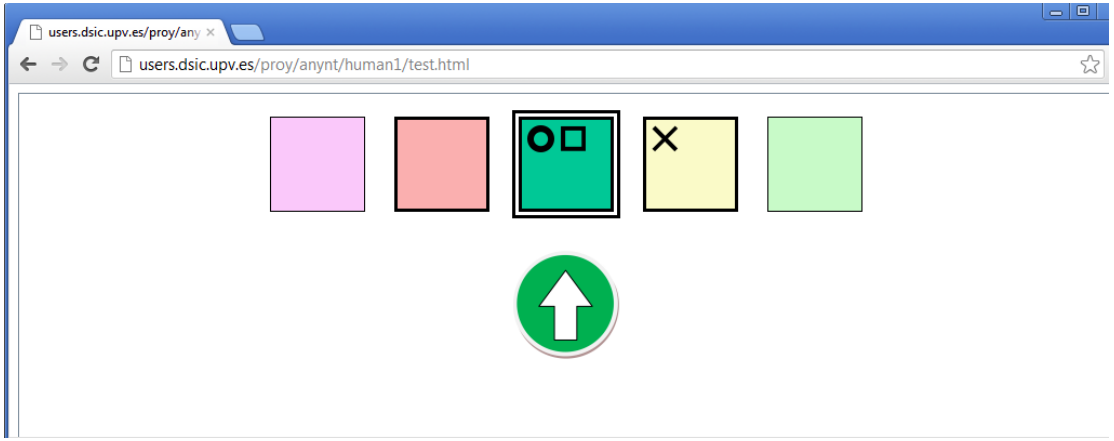
- Anytime universal test (Hernandez-Orallo & Dowe 2010, Artificial Intelligence):
  - The class of environments is carefully selected to be discriminative.
  - Environments are randomly sampled from that class.
    - Starts with very simple environments.
    - Complexity of the environments adapts to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

UNIVERSAL PSYCHOMETRICS

- The anYnt project (2009-2011):

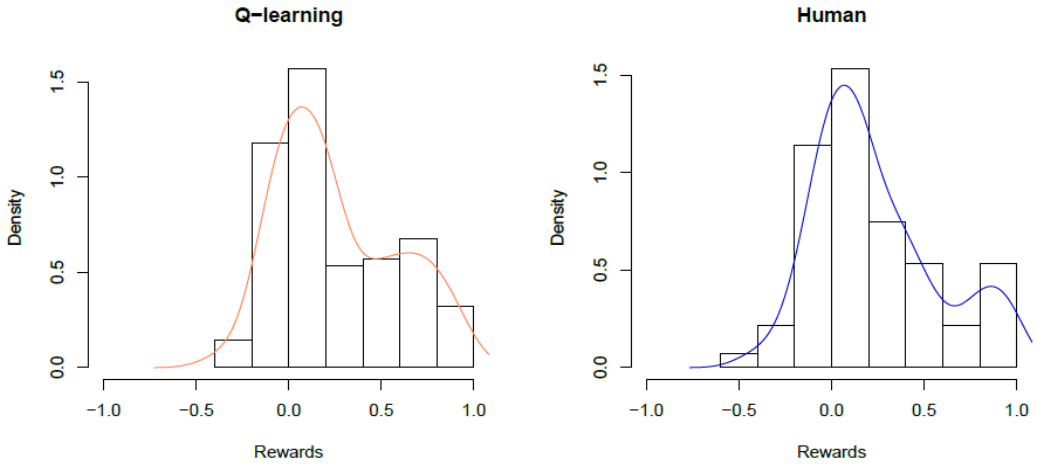
  http://users.dsic.upv.es/proy/anynt/

  - 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.



UNIVERSAL PSYCHOMETRICS

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



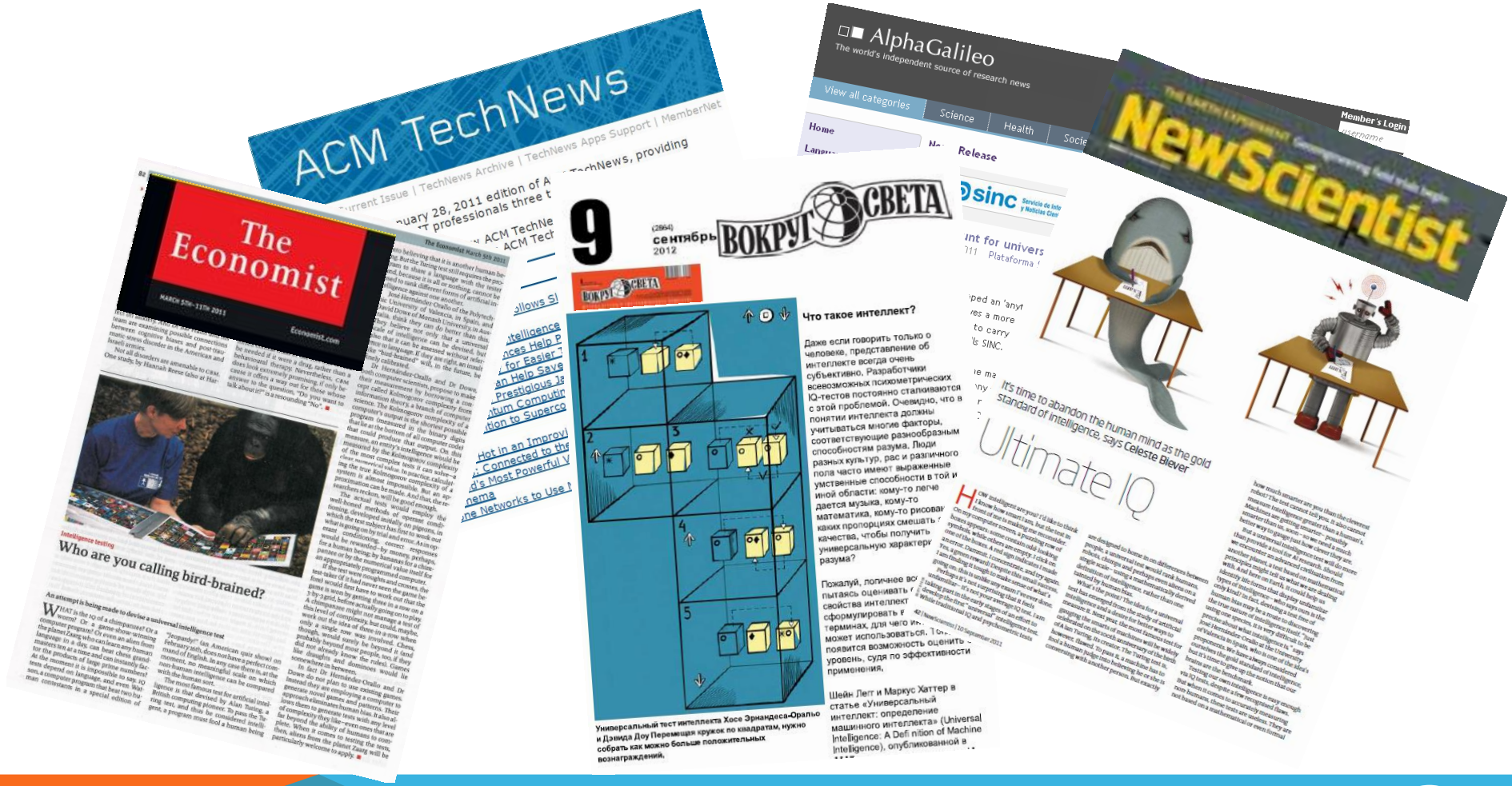
  - Impressions:
    - The test is useful to compare and scale systems of the same type.
    - The results do not reflect the actual differences between humans and Q-learning.

UNIVERSAL PSYCHOMETRICS

- How should this *Popperian refutation* 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).
  - Difficulty not used for Φ (Q-learning relentless for easy problems)
- Are universal tests impossible?
  - All the above issues should be explored before dismissing this idea.

UNIVERSAL PSYCHOMETRICS

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



UNIVERSAL PSYCHOMETRICS

- *Something went very wrong here…*

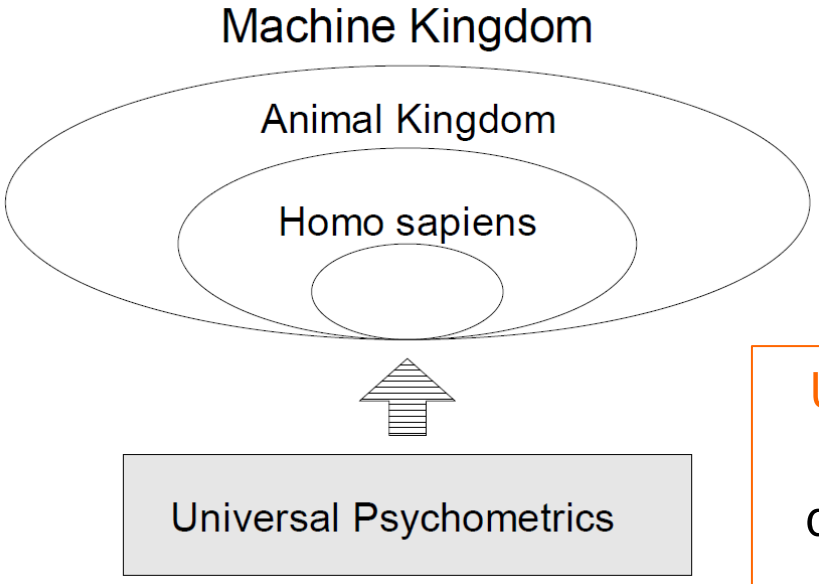
UNIVERSAL PSYCHOMETRICS

- 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 (Hernández-Orallo et al, 2014, Cog Sci Res) is the analysis and development of measurement techniques and tools for the evaluation of cognitive abilities of subjects in the machine kingdom.

UNIVERSAL PSYCHOMETRICS

- Elements:
  - Subjects: physically computable (resource-bounded) interactive systems.
  - Cognitive task: physically computable interactive systems with a score function.
  - Cognitive ability (or task class): set of cognitive tasks.
    - The separation between task-specific and ability-specific evaluation becomes a progressive thing, depending on the generality of the class.
  - 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.

UNIVERSAL PSYCHOMETRICS

- Interdisciplinary approach:
  - Integrates many elements found in psychometrics and comparative cognition
    - Overhauls them with the theory of computation and AIT.
  - Uses a universal notion of intelligence.
    - Tests can be universal or not, depending on the application.
  - Tests for different kinds of individuals are most informative
    - They can falsify attempts for intelligence tests.
  - Comparison and standardisation of difficulty
    - Theoretical notions with observed notions of difficulty in different (biological) populations (humans or animals).

UNIVERSAL PSYCHOMETRICS

- Principled views about the evaluation of:
  - *General intelligence vs particular abilities*:
    - *g factor:* is it a biological phenomenon of evolution or a universal finding of intelligence?
  - *Potential intelligence*:
    - a UTM can emulate any other machine.
  - *Social intelligence*:
    - confronting agents with other agents, Darwin-Wallace distribution.
  - *Collective intelligence*:
    - emergent abilities in terms of the individual abilities of the group.
      - Determine qualities for machines in groups as facilitators, mediators, negotiators, etc.

DISCUSSION

- Universal psychometrics can be seen as a cybernetic approach:
  - Embracing "*the animal and the machine*" (Wiener 1948):
  - Towards further generalisation.

- And as in general systems theory.
  - *"There exist models, principles, and laws that apply to generalized systems or their subclasses, irrespective of their particular kind, the nature of their component elements, and the relationships or "forces" between them. It seems legitimate to ask for a theory, not of systems of a more or less special kind, but of universal principles applying to systems in general." (Bertalanffy 1968)*

An interdisciplinary challenge

- With a formal foundation on mathematics and computation.

DISCUSSION

- Is “intelligent” system evaluation crucial now?
  - New “intelligent” systems abound:
    - As assistants and peers in work teams, social networks, organisations, etc.
    - Many decisions are being delegated to these systems:
      - Business intelligence, big data, data science, cognitive computing (Watson), etc.

- Questions:
  - What are their cognitive abilities?
  - Which tasks are they able to do?
  - How do they work in teams?
  - Are they unreliable?

We need tools to evaluate (or even certify) these systems

CONCLUSIONS

- Two views of “intelligent” system evaluation
  - Task-oriented evaluation
    - Still a huge margin of improvement in the way AI systems are evaluated.
    - The key issues are *M* and *p*, and distinguishing the definition of the problem class from an effective sampling procedure (testing procedure).
  - Ability-oriented evaluation
    - The notion of ability is more elusive than the notion of task.
    - Scattered efforts in AI, psychometrics, AIT and comparative cognition:
      - Universal psychometrics as a unified view for evaluation of cognitive abilities.
- More a matter of degree as sets of tasks become wider.

CONCLUSIONS

- Measuring intelligence is a key ingredient for understanding what intelligence is (and, of course, to devise intelligent artefacts).
- Increasing need for system evaluation:
  - Plethora of bots, robots, artificial agents, avatars, control systems, 'animats', hybrids, collectives, etc., systems that develop and change with time.
  - Crucial for the *technological singularity* once (and if) achieved.
- A challenging problem…

QUESTIONS?

* A paper closely related of this presentation, including full coverage of topics and references can be found at: http://arxiv.org/abs/1408.6908

- *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."
>
> Stanisław Lem, "The Futurological Congress (1971)"