CHARACTERISING THE FUTURE OF INTELLIGENCE THROUGH MEASUREMENT*

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

* Based on parts of the book: “The Measure of All Minds”: http://allminds.org

Seminar at CSER, Cambridge, 19 August 2016
“Greatest accuracy, at the frontiers of science, requires greatest effort, and probably the most expensive or complicated of measurement instruments and procedures”
(David Hand, 2004).
OUTLINE

- Extended Nature
- The Evaluation Discordance
  - Psychometrics (Humans)
  - Comparative Cognition (Animals)
  - AI Evaluation (Non-biological Machines)
  - Mismatch (Cross-refutation)
- The Algorithmic Confluence
  - Why AIT? (Compression and Induction)
  - What IQ Tests Measure (the C-tests)
  - Situated Tests (Range of Environments)
  - Solutional Approach (Levin’s Universal Search)
- Implications
  - Arranging Tasks and Abilities (The Structure of the Ars Generalis)
  - Ratiocentric Perspective (The Kingdom of Ends)
- Conclusions
“There is a label on a cage that states simply, ‘This machine is new to science’. Inside the cage there sits a small dustbot. It has bad temper. No bad-tempered dustbot has ever been found. Nothing is known about it. It has no name. For the mechanist it presents an immediate challenge. What has made it unique? How does it differ from the other dustbots already known and described?”*


- Sets a non-anthropocentric, unprejudiced perspective.
EXTENDED NATURE

- Computers:
  - AI or AGI systems, (cognitive) robots, bots, ...
- Cognitively-enhanced organisms
  - Cyborgs, technology-enhanced humans (cognitive prosthetics).
- Biologically-enhanced computers:
  - Human computation
- (Hybrid) collectives
  - Virtual social networks, crowdsourcing
- Minimal or rare cognition
  - Plants, bacteria, artificial life, ...
Universal psychometrics is the analysis and development of measurement tools for the evaluation of behavioural features in the machine kingdom, including cognitive abilities and personality traits.
A behavioural feature is an abstract property, characteristic or construct about the elements of the machine kingdom that can be inferred by observation and interaction.
EXTENDED NATURE

- Why inferred from behaviour *(black-box)*?
  - Evaluate systems that we have not designed or are unpredictable:
    - Biological, AI and hybrid systems: complex, stochastic, faulty, ...
  - In many virtual ecosystems

  **We want the ground truth, which is ultimately behaviour!**

- Why abstract features?
  - Fine-grained models from behaviour? Guess my next action?
    - Not even with white-box approaches: *Rice’s theorem*
  - Stable predictive/descriptive features? Am I extroverted? Intelligent?
    - This is one of the key questions of behavioural sciences

  **A psychometric profile: measured behavioural features.**
THE EVALUATION DISCORDANCE

- There is plenty of evaluation techniques in different disciplines!
  - Human evaluation: psychometrics
  - (Non-human) animals: comparative psychology
  - AI systems: AI competitions and benchmarks

- The three follow some common principles of measurement:
  - a representational approach or a pragmatic approach:

- A behavioural feature as an aggregation of task results for a subject:

\[ \Psi(\pi, M, p) \triangleq \sum_{\mu \in M} p(\mu) \cdot R(\pi, \mu) \]

Performance as agglomerated result
How are tests conceived?

- **Personality traits:** lexical hypothesis (Galton, 1884),
- **Cognitive abilities:** culture-fair tasks, no “idiots savants”.

### How Accurately Can You Describe Yourself?

Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. So that you can describe yourself in an honest manner, your responses will be kept in absolute confidence. Indicate for each statement whether it is Very Inaccurate (VI), Moderately Inaccurate (MI), Neither Accurate Nor Inaccurate (NN), Moderately Accurate (MA), or Very Accurate (VA), as a description of you.

<table>
<thead>
<tr>
<th>Statement</th>
<th>VI</th>
<th>MI</th>
<th>NN</th>
<th>MA</th>
<th>VA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Am the life of the party</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Feel little concern for others</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Am always prepared</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Get stressed out easily</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Have a rich vocabulary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Don’t talk a lot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Am interested in people</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Leave my belongings around</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Am relaxed most of the time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Have difficulty understanding abstract ideas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- carpenter: wood
- sun: planet
- teacher: chalk
- mason: ________
- earth: ________
- soldier: ________
**How is difficulty identified?**

- **Item Response Theory:**
  - Logistic models: difficulty $b$, discrimination $a$ and guessing $c$.

---

**3PL-model**

$$p(\theta) = c + \frac{1 - c}{1 + e^{-a(\theta-b)}}$$

**Item Parameters**

- Subject Proficiencies $\theta$

**MLE:** e.g. Birnbaum method

**Result matrix**

<table>
<thead>
<tr>
<th>Tests</th>
<th>Subjects</th>
</tr>
</thead>
</table>

**Proficiency as achievable difficulty**
THE EVALUATION DISCORDANCE: Psychometrics

- But how are features identified?
  - **Personality traits:** clustering (big five).
  - **Cognitive abilities:** primary abilities, g factor, hierarchical models.

<table>
<thead>
<tr>
<th>Tests</th>
<th>Subjects</th>
<th>Result matrix</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Latent factors</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Tests</th>
</tr>
</thead>
</table>

- Psychometrics relies on a data-driven **populational** approach.
  - Groups are a combination of **nature** and **nurture**.
  - Changing with time (**Flynn effect**).

- **IQ score versus g score:** IQ scores strongly depend on the composition of the agglomeration of tasks in an IQ test whereas **g scores** are less dependent.

- **IQ score versus g score:** IQ scores strongly depend on the composition of the agglomeration of tasks in an IQ test whereas **g scores** are less dependent.

Theories of intelligence

Catell-Horn-Carroll hierarchical model
THE EVALUATION DISCORDANCE: COMPARATIVE COGNITION

- Evaluating animals is far more complex!
- Easy to misinterpret results from anecdotal experience or specific tasks:

- Contaminated by anthropocentrism:

Lloyd Morgan’s Canon: “In no case is an animal activity to be interpreted in terms of higher psychological processes if it can be fairly interpreted in terms of processes which stand lower in the scale of psychological evolution and development” (Morgan, 1903, p. 59).

Images from BBC One documentary: “Super-smart animal”: http://www.bbc.co.uk/programmes/b01by613
THE EVALUATION DISCORDANCE: COMPARATIVE COGNITION

- More systematic animal evaluation:
  - Use of rewards and sophisticated interfaces (testing apparatus).

- Becoming less anthropocentric: humans as a special case.
- Inter-species and intra-species analysis: controversial findings: e.g., animal g (intra-species) and G (inter-species).

Anecdotal, qualitative research shifting to more systematic, quantitative studies for both personality traits and cognitive abilities in animals.
"[AI is] the science of making machines do things that would require intelligence if done by [humans]."
Marvin Minsky (1968).

- They can do the “things” (tasks) without featuring intelligence.
- Once the task is solved (“superhuman”), it is no longer an AI problem (“AI effect”)
  - AI would have progressed very significantly (see, e.g., Nilsson, 2009, chap. 32, or Bostrom, 2014, Table 1, pp. 12–13).
- But the machine kingdom is now full of idiots savants.
THE EVALUATION DISCORDANCE: AI EVALUATION

- Specific (task-oriented) AI systems

- Prediction and estimation
- Machine translation, information retrieval, summarisation
- Computer vision, speech recognition, etc.
- Expert systems
- Driverless vehicles
- Planning and scheduling
- Automated deduction
- Game playing

Warning! Intelligence NOT included.

All images from wikicommons
THE EVALUATION DISCORDANCE: AI EVALUATION

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 → 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
- ...
THE EVALUATION DISCORDANCE: AI EVALUATION

How can we evaluate general-purpose systems?

- Cognitive robots
- Pets, animats, and other artificial companions
- Agents, avatars, chatbots
- Web-bots, Smartbots, Security bots...
- Smart environments

**Warning!** Some intelligence MAY BE included.
THE EVALUATION DISCORDANCE: AI EVALUATION

- “Mythical Turing Test” (Sloman, 2014) and its myriad variants.
- Mythical human-level machine intelligence
  - A red herring for general-purpose AI!

- General-purpose AI evaluation (AGI systems, no initial functionality!):
  - Trend: video games and sandboxes: General Video Game AI Competition (VGDL), OpenAI Gym, Microsoft’s Malmo

- Many open questions:
  - Task analysis, their similarities, difficulties.....
  - Abilities: be conceptualised and identified.
    - Ability-oriented (or feature-oriented) evaluation

EGPAI2016 - Evaluating General-Purpose AI
Many differences in the three main disciplines

<table>
<thead>
<tr>
<th></th>
<th>Human Psychometrics</th>
<th>Comparative Psychology</th>
<th>Artificial Intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose</td>
<td>diagnosis, selection, education, science</td>
<td>science</td>
<td>engineering</td>
</tr>
<tr>
<td>Subjects</td>
<td>one to many</td>
<td>usually few</td>
<td>usually few</td>
</tr>
<tr>
<td>Normed</td>
<td>population</td>
<td>unusual</td>
<td>no</td>
</tr>
<tr>
<td>Tests</td>
<td>standard battery</td>
<td>single experiment</td>
<td>benchmark</td>
</tr>
<tr>
<td>Interfaces</td>
<td>pen &amp; paper, computerised</td>
<td>specialised</td>
<td>inputs &amp; outputs, robotic</td>
</tr>
<tr>
<td>Motivation</td>
<td>instructions</td>
<td>rewards &amp; penalties</td>
<td>hardwired</td>
</tr>
<tr>
<td>Reliability</td>
<td>usually high</td>
<td>variable</td>
<td>usually low</td>
</tr>
<tr>
<td>Overfitting</td>
<td>usually low</td>
<td>variable</td>
<td>usually high</td>
</tr>
<tr>
<td>Situated</td>
<td>unusual</td>
<td>common</td>
<td>common</td>
</tr>
</tbody>
</table>
THE EVALUATION DISCORDANCE: MISMATCH

- What happens when tests are used across disciplines?
  - Let’s use psychometric tests in AI!
    - In 2003, Sanghi & Dowe implemented a small program scoring relatively well on many IQ tests.

\[\text{This made the point unequivocally: this program is not intelligent}\]

- This has not been a deterrent!
  - Psychometric AI (Bringsjord and Schmimanski 2003):
    - An “agent is intelligent if and only if it excels at all established, validated tests of intelligence”.
  - Detterman, editor of the Intelligence Journal, posed “A challenge to Watson” (Detterman 2011)
    - 2nd level to “be truly intelligent”: tests not seen beforehand.
      - “IQ tests are not for machines, yet” (Dowe & Hernandez-Orallo 2012)
THE EVALUATION DISCORDANCE: MISMATCH

- What about developmental tests (or tests for children)?
  - Developmental robotics: battery of tests (Sinapov, Stoytchev, Schenk 2010-2013)
  - Cognitive architectures:
    - Newell “test” (Anderson and Lebiere 2003)
    - “Cognitive Decathlon” (Mueller 2007).
  - AGI: high-level competency areas (Adams et al. 2012), task breadth (Goertzel et al 2009, Rohrer 2010), robot preschool (Goertzel and Bugaj 2009).

<table>
<thead>
<tr>
<th>Category</th>
<th>Level</th>
<th>PEBL</th>
<th>CHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invariant Object Identification</td>
<td>Yes</td>
<td>Gv</td>
<td></td>
</tr>
<tr>
<td>Object ID: Size Discrimination</td>
<td>Yes</td>
<td>Gv</td>
<td></td>
</tr>
<tr>
<td>Vision</td>
<td>Object ID: With Rotation</td>
<td>Yes</td>
<td>Gv</td>
</tr>
<tr>
<td></td>
<td>Object ID: Relations</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Visual Action/Event Recognition</td>
<td>No</td>
<td>Gv</td>
</tr>
<tr>
<td>Search</td>
<td>Simple Navigation</td>
<td>Yes</td>
<td>Gv</td>
</tr>
<tr>
<td></td>
<td>Visual Search</td>
<td>Yes</td>
<td>Gv</td>
</tr>
<tr>
<td></td>
<td>Travelling Salesman Problem</td>
<td>Yes</td>
<td>Gv</td>
</tr>
<tr>
<td></td>
<td>Embodied Search</td>
<td>No</td>
<td>Gv</td>
</tr>
<tr>
<td></td>
<td>Reinforcement Learning</td>
<td>Yes</td>
<td>Gv</td>
</tr>
<tr>
<td>Knowledge Learning</td>
<td>Episodic Recognition Memory</td>
<td>No</td>
<td>Gv</td>
</tr>
<tr>
<td></td>
<td>Semantic Memory/Categorization</td>
<td>No</td>
<td>Gl</td>
</tr>
<tr>
<td>Language and Concept Learning</td>
<td>Object-Noun Mapping</td>
<td>No</td>
<td>Gc</td>
</tr>
<tr>
<td></td>
<td>Property-Adjective</td>
<td>No</td>
<td>Gc</td>
</tr>
<tr>
<td></td>
<td>Relation-Preposition</td>
<td>No</td>
<td>Gc</td>
</tr>
<tr>
<td></td>
<td>Action-Verb</td>
<td>No</td>
<td>Gc</td>
</tr>
<tr>
<td></td>
<td>Relational Verb-Action</td>
<td>No</td>
<td>Gc</td>
</tr>
<tr>
<td>Simple Motor Control</td>
<td>Eye Movements</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Aimed Manual Movements</td>
<td>Yes</td>
<td>-</td>
</tr>
</tbody>
</table>
THE EVALUATION DISCORDANCE: MISMATCH

- Test from one group only valid and reliable for the original group.
- An opportunity:
  
  ![Diagram showing human evaluation, animal evaluation, and AI evaluation]

  The machine kingdom is a vast source for refutation.

  - How universal can a test be?: working for the whole machine kingdom!
  - With populations we were able to derive representative tasks, arrange them into abilities and derive difficulty notions
    - Can we define a population of machines?
      - A discrete distribution over computable machines?
      - According to state of the art in AI?

We need a different foundation
THE ALGORITHMIC CONFLUENCE: Why AIT?

- Algorithmic Information Theory
  - Kolmogorov complexity, $K_U(s)$: shortest program for $U$ outputting $s$.
  - Algorithmic probability (universal distribution), $p_U(s)$: the probability of objects as outputs of a UTM $U$ fed by a fair coin. Also: $p_U(s) = 2^{-K_U(s)}$

Compression and inductive inference (and learning): (Occam’s razor, Solomonoff’s prediction, MML/MDL, NFL...).

- Invariance theorem: $K_{U_1}(s)$ and $K_{U_2}(s)$ only differ by a constant (ind. $s$).
- $K(s)$ is incomputable, but approximations exist (Levin’s $K_t$).

$$Kt_U(x) \triangleq \min_{p : U(p)=x} LS(p)$$

with

$$LS(p) \triangleq L(p) + \log S(p)$$

- Any inversion problem can be solved optimally by Levin’s search (except for a multiplicative constant), with upper bound $S(p)2^{L(p)}$.

$K_t$ linked to Levin’s universal search: “universal heuristics” (Levin 2013).
“Beyond the Turing Test”...

“Intelligence” definition and test (C-test) based on algorithmic information theory (Hernandez-Orallo 1998-2000).

- Letter series common in cognitive tests (Thurstone).
- Here generated from a TM with properties (projectibility, stability, ...).

\[
\begin{align*}
  k = 9 & : a, d, g, j, \ldots & \text{Answer: } m \\
  k = 12 & : a, a, z, c, y, e, x, \ldots & \text{Answer: } g \\
  k = 14 & : c, a, b, d, b, c, c, e, c, d, \ldots & \text{Answer: } d
\end{align*}
\]

- Their difficulty is calculated by Kt
  - Linked with Levin’s universal search.
THE ALGORITHMIC CONFLUENCE: WHAT IQ TESTS MEASURE

- Metric derived by slicing by difficulty $h$ ($Kt$) and:

$$I(\pi) \triangleq \sum_{h=1}^{H} h^e \sum_{i=1}^{N} \frac{1}{N} \text{Hit}(\pi, x_i, h)$$

- This is IQ-test re-engineering! IQ tests formal and well grounded.
  - Intelligence no longer “what intelligence tests measure” (Boring, 1923).
  - Clues about what IQ tests really measure? Inductive inference.

But remember Sanghi and Dowe 2003!

- Extensions suggested (Hernández-Orallo 2000b, PerMIS):
  - “use of rewards and penalties”, “other factors”.

Human performance correlated with the difficulty ($h$) of each exercise.
- Intelligence as **performance in a range of worlds**.
  - **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 aggregate, following the usual average-case evaluation:

\[
\Psi(\pi, M, p) \triangleq \sum_{\mu \in M} p(\mu) \cdot R(\pi, \mu)
\]
THE ALGORITHMIC CONFLUENCE: SITUATED TESTS

- In an alternately-synchronous RL setting choosing \( p(\mu) = 2^{-K(\mu)} \) we get the so-called “Universal Intelligence” (Legg and Hutter 2007).
  - Proper formalisation of a range of environments.
  - Problems (pointed out by many: Hibbard 2009, Hernandez-Orallo & Dowe 2010):
    - The mass of the probability measure goes to a few environments.
    - \( M \) or the probability distribution is not computable.
    - Most environments are not really discriminating (hells/heavens).
    - There are two infinite sums (environments and interactions).
    - Time/speed is not considered for the environment or agent.
  - Many now acknowledged (see Leike & Hutter 2015).

Legg and Hutter’s measure is “relative”, a schema for tasks, a meta-definition instantiated by a particular universal distribution.
THE ALGORITHMIC CONFLUENCE: SOLUTIONAL APPROACH

- Recovering the slicing by difficulty in the C-test.
  - We need a general measure of difficulty $h(\mu)$.
    - Adapt Levin’s universal search as “policy search effort”.
    - A threshold for acceptable policies:
      $$A^{[\varepsilon, \nu]}(\mu) \Delta \{ \pi : R^{[\varepsilon, \nu]}(\pi, \mu) \geq 1 - \varepsilon \}$$
      $$A(\pi, \mu) \Delta 1 \text{ if } \pi \in A(\mu) \text{ and } 0 \text{ otherwise}$$
  - Not about the resources/size of the task but the solution.

<table>
<thead>
<tr>
<th>Feature of $\pi$</th>
<th>Difficulty</th>
<th>Notation</th>
<th>Depends on</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected response</td>
<td>-</td>
<td>$R(\pi, \mu)$</td>
<td>$\tau, \nu$</td>
</tr>
<tr>
<td>Variance of response</td>
<td>-</td>
<td>$\text{Var}[R(\pi, \mu)]$</td>
<td>$\tau, \nu$</td>
</tr>
<tr>
<td>Information (size)</td>
<td>Transmission</td>
<td>$L(\pi)$</td>
<td>-</td>
</tr>
<tr>
<td>Execution steps</td>
<td>Demonstration</td>
<td>$S(\pi, \mu)$</td>
<td>$\tau, \nu$</td>
</tr>
<tr>
<td>Finding effort steps</td>
<td>Finding</td>
<td>$LS(\pi, \mu)$</td>
<td>$L, S, \varepsilon$</td>
</tr>
<tr>
<td>Verification steps</td>
<td>Verifying</td>
<td>$W(\pi, \mu)$</td>
<td>$S, \text{Var}[R(\pi, \mu)], \varepsilon, \delta$</td>
</tr>
<tr>
<td>Total effort steps</td>
<td>Search</td>
<td>$F(\pi, \mu)$</td>
<td>$L, W$</td>
</tr>
</tbody>
</table>

$$h^{[\varepsilon, \delta, \nu]}(\mu) \Delta \min_{\pi \in A^{[\varepsilon, \nu]}(\mu)} F^{[\varepsilon, \delta, \nu]}(\pi, \mu)$$
We can slice the performance measure:

$$\Psi_h(\pi, M, p_M) \triangleq \sum_{\mu \in M, h(\mu) = h} p_M(\mu|h) \cdot A(\pi, \mu)$$

And use it in the general equation.

$$\Psi(\pi, M, p_M) = \sum_{h=0}^{\infty} p_H(h) \Psi_h(\pi, M, p_M)$$

From Bayes, $p(h)$ and $p(\mu|h)$ can reconstruct any original $p(\mu)$.

$$\Psi(\pi, M, p_M) \triangleq \sum_{\mu \in M} p_M(\mu) \cdot A(\pi, \mu)$$

But this is not our goal!
The algorithmic confluence: Solutional Approach

- The psychometrician sieve:
  - Three approaches:

1. Direct: Items derive from the representational distribution
   - $p_H(h)$ [universal]
   - $p_M(\mu)$ [universal]
   - $\mu$

2. Indirect: Items derive from difficulty.
   - $p_H(h)$ [uniform]
   - $p_M(\mu|h)$ [universal]
   - $\mu$

3. Indirect through policies: Items derive from policies, policies from difficulty.
   - $p_H(h)$ [uniform]
   - $p_H(\pi|h)$ [uniform]
   - $\pi$
   - $p_M(\mu|\pi)$ [universal]
   - $\mu$

- With the choices in brackets, they are NOT equivalent.
THE ALGORITHMIC CONFLUENCE: SOLUTIONAL APPROACH

- This suggests a different view of “general intelligence”, different from an agglomerative task-general intelligence (universal intelligence):
  - Policy-general intelligence: aggregate by difficulty (e.g., bounded uniform) and for each difficulty look for diversity.

  Ability to find, integrate and emulate a diverse range of successful policies.

- Connected to the task-independence of the $g$ factor.

<table>
<thead>
<tr>
<th></th>
<th>Human population</th>
<th>Universal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agglomerative</td>
<td>IQ</td>
<td>Task-general intelligence</td>
</tr>
<tr>
<td>Solutional</td>
<td>$g$ factor</td>
<td>Policy-general intelligence</td>
</tr>
</tbody>
</table>

Raises a fascinating question:
Is there a universal $g$ factor?
Guiding AI progress and characterising different kinds of intelligence depends on the intermediate levels between tasks and abilities:

- Task breadth? Arrange abilities?
  - Hierarchically or like Guttman’s model

- Intrinsic notion of similarity between tasks?

- Makes sense again if sliced by difficulty:
**IMPLICATIONS: ARRANGING TASKS AND ABILITIES**

- Example (ECA rules as tasks).
  - Task description is not used. No population is used either.
  - The best solutions are used instead and compared.
- Using **similarity as difficulty increase** (18 rules of difficulty 8):

![Dendrogram using complete linkage](image-url)
**IMPLICATIONS: RATIOCENTRIC PERSPECTIVE**

- The evaluation of cognitive abilities crucial for:
  - **Cognitive development**: difficulty relative to previous knowledge.
  - **Potential abilities**: cognitive enhancement.
  - **Social skills**: how do they relate to general intelligence?
  - **Communication and language**: skills expected in the peer?
  - **Collective and hybrid systems**: psychometric profile of group/individual.
  - **Superintelligence**: hypotheses can be falsified.

- Several facets:

<table>
<thead>
<tr>
<th>Policy acquisition</th>
<th>Non-incremental</th>
<th>Incremental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search</td>
<td>Developmental</td>
<td>Developmental and social</td>
</tr>
<tr>
<td>Demonstration</td>
<td>Social</td>
<td>Developmental and social</td>
</tr>
<tr>
<td>Transmission</td>
<td>Social and verbal</td>
<td>Developmental, social and verbal</td>
</tr>
</tbody>
</table>
IMPLICATIONS: RATIOCENTRIC PERSPECTIVE

- A fundamental approach for the analysis of “the future of intelligence”
  - The ‘extant personity’: “kingdom of ends”, minds or moral subjects.
    - By challenging the equation "human=person":

The ratiocentric (not the sentiocentric) view to moral agency and patiency, based on potential abilities and self-modification.

Great responsibility on evaluation:
- Delicate issues about the nature and rights of individuals.
- Non-monolithic and non-anthropocentric view of AI risks and super-intelligence, linked to computational resources.
CONCLUSIONS

- Measuring intelligence is a key ingredient to understand what intelligence is (and, of course, to devise intelligent artefacts).

- Increasing need for system evaluation:
  - Plethora of new systems: AI, hybrids, collectives, etc., developing.
  - Crucial to assess ability profiles unlike and beyond human profiles.

- A challenging shift:
  - From a populational to a universal perspective,
  - From agglomerative to solutional approaches,
  - From a task-general intelligence to a policy-general intelligence.

Universal psychometrics as a unified view of the evaluation of behavioural features of any kind of “mind”.
THANK YOU!

http://www.allminds.org

Major questions addressed in the book:
1. How can behavioural features be measured in the machine kingdom?
2. How universal and adaptive can a behavioural test be?
3. Are IQ tests valid for any machine and what do they measure?
4. How can abilities be identified, through task breadth or similarity?
5. Is intelligence one or many, objective or subjective?
6. Can we formalise task difficulty without particular populations?
7. Is there a common component for all abilities, a universal g factor?
8. Can general intelligence be independent of the task distribution?
9. Will scales fully range from minimal cognition to supercapacities?
10. Can intelligence be defined solely with computational principles?
11. Should collectives and hybrids be evaluated unlike their individuals?
12. Is intelligence useful for adaptive testing and self-assessment?
13. Can social abilities be measured through situated tests?
14. What kind of ability can capture communication and language skills?
15. How can potential abilities be measured and how do they develop?
16. Can any universal machine and human become arbitrarily intelligent?
17. How can cognitive abilities and personality traits develop in general?
18. Are behavioural features sufficient to characterise personhood?
19. What are the limits and consequences of supercapacities?
20. Can human, animal and AI evaluation benefit from an integration?
THANK YOU!
MANY (RESEARCH) POSSIBILITIES

Examples:

- Application of techniques from psychometrics and comparative cognition to AI.
  - LLTM for difficulty estimation in Malmo’s tasks, AAAI 2017?.

- Application of AI theory and AIT to questions in the social sciences:

- Refutation of the universality of IQ tests:

- Feature-oriented vs task-oriented view of the automatisation of work.

- Universal g factor?