Towards an Artificially Intelligent System

Evaluating the Intelligence of an Artificial System

Ondřej Vadinský

Department of Information and Knowledge Engineering
University of Economics Prague

December 2016

Overview

Introduction

- *Motivation*: **Artificial intelligence** (AI) in all aspects **comparable to human intelligence**.
- *PHD Research Topic* – **Analysis of Presumptions of Intelligence of Computer Systems**.
- Multidisciplinary approach: *AI*, *cognitive science*, and *philosophy*.
- *The talk:*
  - **How to recognize that an artificial system is intelligent?**
  - *Overview of approaches defining, and evaluating intelligence.*
  - *Replicating experiments with AIQ test, and how to improve it.*

Philosophical Presumptions of Intelligence

- Descartes:
  - **Universality of thought**,
  - Ability of **rational speech**.
- Turing – *imitation game*, aka *Turing test:*
  - **Human language communication**.
- Harnad – *Total Turing test:*
  - **Human intelligent behavior**.
- Schweizer – *Trully Total Turing test:*
  - Evolution of **intelligent behavior of a species**.
- Searle – *Chinese room argument*:
  - **Meaning**, **understanding**, **intentionality**.
- Dennett:
  - Intelligence and **consciousness**.

Cognitive Presumptions of Intelligence

- De Mey – *Cognitive paradigm*:
  - Representation, aka **world model**(s).
  - **Interaction with the world** (perception, expectation, action).
  - **Formation of knowledge** from implicit to explicit.
- Sun – *hierarchical explanation of cognition*:
  - **Levels of abstraction** (social, psychological, componential, physiological),
  - **Set of constraints** from upper and lover levels.
- *Cognitive architectures* – **domain-generic computational models of cognition**:
  - Sun – CLARION,
  - Anderson – ACT-R,
  - Thomsen – Ouroboros.

Strong, Weak, Specific and General AI

- Searle:
  - *Strong AI* – **AI** as an **explanation** and **duplication of the mind**,
  - *Weak AI* – **AI** as a **tool for modeling** and **simulation**.
- Goertzel:
  - *Specific AI* – **solving a certain task** or a limited set,
  - *General AI* (AGI) – **solving a broad set of task**.

Psychometric AI

- Bringsjord and Schimanski (2003) – *psychometric AI* (PAI):
  - Uses **psychometric definitions of intelligence**.
  - Psychometrics – **measurement of intelligence** in humans using tests.
  - AI should focus on "building information-processing entities capable of at least **solid performance on all established, validated tests of intelligence and mental ability**."
- Besold et al.:
  - **Directly using human intelligence tests is problematic** (necessity, sufficiency).
  - Generalization and improvement of tests is needed.

Universal Intelligence Definition

- Legg and Hutter (2007) – *Universal intelligence* (UI):
  - AI needs a precise and **formal definition of intelligence**.
  - Abstraction of existing psychological definitions of intelligence.
  - **"Intelligence measures an agent's ability to achieve goals in a wide range of environments."**

$$\Upsilon(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_\mu^\pi, \text{ where } \quad V_\mu^\pi := \mathrm{E}\left(\sum_{i=1}^{\infty} r_i\right) \leq 1$$

    - *Agent–environment interaction* (actions, observations, rewards).
    - *All environments weighted by Kolmogorov complexity* (Occam's razor).
    - **Maximizing expected future total rewards given past interactions.**
  - Enables ordering of performance of agents.
  - To achieve high UI a true generality is needed.
  - **Not anthropocentric**, nor culturally biased.

Pragmatic General Intelligence Definition

- Goertzel (2010) – a critique of Universal Intelligence:
  - *Pragmatic General Intelligence*:
    - **Agents choosing their goals**, and **agents being adapted to certain environments**:

      "Intelligence is achieving complex goals in complex environments."

$$\Pi(\pi) \equiv \sum_{\mu \in E, g \in G, T} \nu(\mu)\gamma(g,\mu)V^{\pi}_{\mu,g,T}, \text{ where } \quad V^{\pi}_{\mu,g,T} \equiv \mathrm{E}\left(\sum_{i=s}^{t} r_g(I_{g,s,i})\right)$$

  - *Efficient Pragmatic General Intelligence*:
    - normalized by **computational resource consumption**.
  - *Intellectual Breadth*:
    - **A fuzzy set of contexts** (goals, environments, time intervals), **relative to which an agent is intelligent**.

Algorithmic Intelligence Quotient Test I

- Legg and Veness (2013) – *Algorithmic Intelligence Quotient* (AIQ):
  - **An approximate test of Universal intelligence**.

$$\hat{\Upsilon}(\pi) := \frac{1}{N}\sum_{i=1}^{N} \hat{V}_{p_i}^{\pi}, \text{ where } p_i \text{ chosen by } \quad M_{\mathcal{U}}(x) := \sum_{p:\mathcal{U}(p)=x*} 2^{-l(p)}$$

    - *Finite sample of environment programs* (shorter are preferred).
    - Solomonoff's *Universal Distribution* in place for Occam's razor.
    - *Limited number and time of agent–environment interactions.*
  - **Open Source prototype implementation:**
    - *set of supplied simple agents*, others through a wrapper,
    - configurable *number of tested environment programs*, and *number of agent–environment interactions*,
    - configurable size of *action and observation space.*

Algorithmic Intelligence Quotient Test II

- *Environment program*:
  - **computes current reward** [-100;+100] **and observation from the interaction sequence**,
  - implemented in **extended BF reference machine**:
    - `+-` increment/decrement work cell symbol,
    - `,.` read from input tape/write to output cell,
    - `<>` move work tape one cell left/right,
    - `[]` start a loop if work cell is non-zero/end loop,
    - `%` write a random symbol to work cell,
    - `#` end program.

      `[+.>]-,,#`

      `+[[[+>]<<.,<%->]%,>-<],+[+>.,.%]+>.%%....<+#`
  - *BF reference machine settings*:
    - default: 5 action/observation symbols, 1 observation/reward cell,
    - configurable action space (nr of action/observation symbols),
    - configurable observation space (nr of observation cells).

Algorithmic Intelligence Quotient Test III

- *Agents in the test:*
  - *random* – random behavior,
  - *freq* – chooses action with highest average reward,
  - Q_0 – basic Q-learning,
  - $Q\lambda$ – Q-learning with eligibility traces,
  - $HLQ\lambda$ – Q-learning with automatic learning rate,
  - *MC-AIXI* – wrapper for Monte Carlo approximation of AIXI (optimal agent according to UI definition).

Replicating Experiments with AIQ Test

- *Legg and Veness 2013:*
  - mainly how to convert the Universal Intelligence definition into a practical AIQ test,
  - a brief mention of **experiments**, **without much detail**,
  - if AIQ test is to be improved, concrete results are needed for comparison.
- *Conducted experiments:*
  - **The default settings** – BF 5.
  - **Varying the action space** – BF 2, BF 5, BF 10, and BF 20.
  - **Varying the observation space** – BF 5,1, BF 5,2, BF 5,3.
- *Common settings:*
  - new environment program samples generated,
  - episode length: 1,000, 3,000, 10,000, 30,000, and 100,000 iterations,
  - 5 configurations of each agent used for all reference machines,
  - for MC-AIXI a parameter sweep conducted in default settings

The Default Settings Experiment

Agent	Differences of AIQ Scores with Confidence Intervals for Episode Length									
	1,000		*3,000*		*10,000*		*30,000*		*100,000*	
freq	− 0.1	± 0.6	0.2	± 0.6	− 0.3	± 0.6	− 0.4	± 0.7	*− 1.0*	*± 0.7*
Q_o	0.0	± 0.6	− 0.2	± 0.6	*− 1.2*	*± 0.6*	*− 1.1*	*± 0.7*	*− 1.1*	*± 0.7*
$Q\lambda$	− 0.3	± 0.6	0.0	± 0.6	*− 0.9*	*± 0.7*	*− 1.4*	*± 0.7*	*− 1.0*	*± 0.7*
$HLQ\lambda$	− 0.5	± 0.6	− 0.3	± 0.7	*− 0.9*	*± 0.6*	*− 1.4*	*± 0.7*	*− 1.1*	*± 0.7*
MC-AIXI	**3.5**	**± 1.8**	*2.5*	*± 1.6*	− 0.1	± 1.9	− 1.6	± 1.7	**− 4.3**	**± 1.8**

Table: Comparison of the best result achieved for each agent at a given EL.

- **In accordance with the original results**.
- There are *some significant (small) differences*, however the *ordering of agents according to maximal AIQ remains the same as original.*
- *For MC-AIXI both higher AIQs (for shorter EL) and lower AIQs (for longer EL) were achieved*, however the original configurations are not known.

Varying the Action Space Experiment

- **Contrary to the original results**.
- *AIQ means differ significantly* for BF 2×BF 20, BF 5×BF 20, and BF 10×BF 20.
- *SD means of AIQ differ significantly* among BF 2, BF 5, BF 10, and BF 20.
- The ordering of agents according to maximal AIQ changes compared to default experiment.
- AIQ scores of agents decline with increasing action space, except for Freq.
- MC-AIXI is especially affected.
- The effects may be due to agents configurations being too fine-tuned for BF 5.

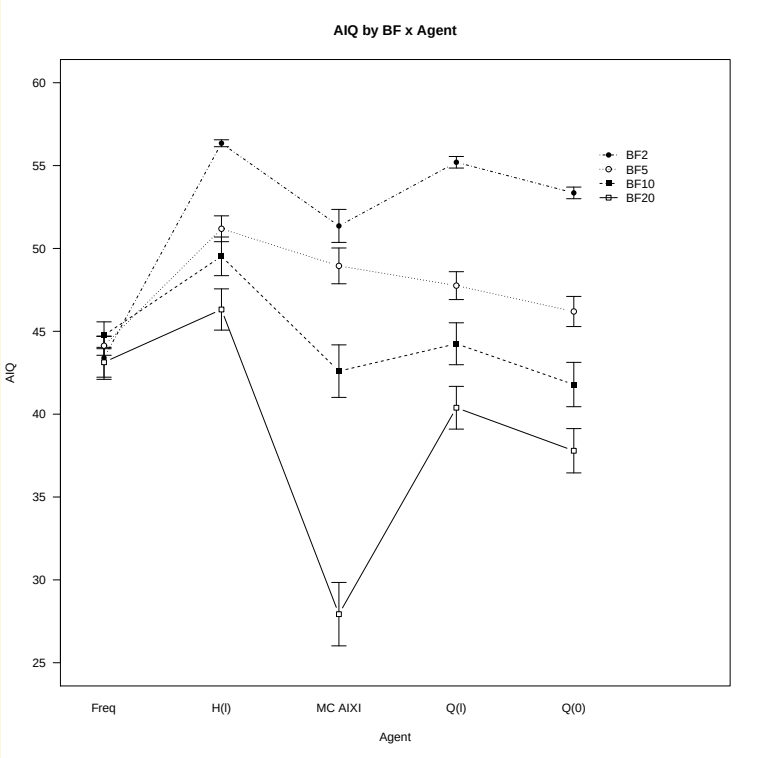

Figure: The impact of changing BF machine on the mean AIQ scores of agents.

Varying the Observation Space Experiment

- **In accordance with the original results**.
- *AIQ means do not differ significantly* among BF 5,1, BF 5,2, and BF 5,3
- *SD means of AIQ do not differ significantly* among BF 5,1, BF 5,2, and BF 5,3
- The ordering of agents according to maximal AIQ does not change compared to default experiment, except from MC-AIXI.
- MC-AIXI is somewhat affected.

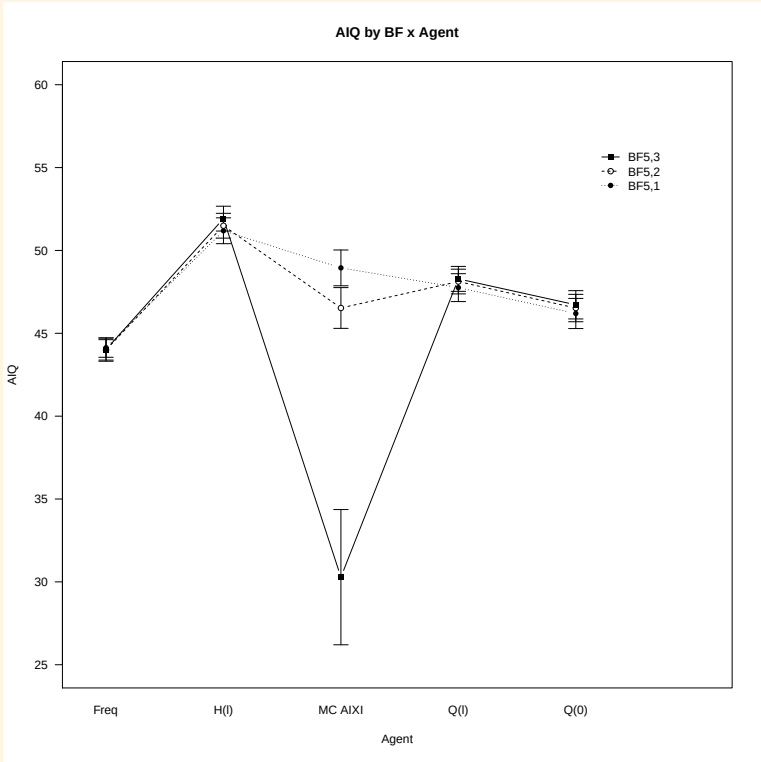

Figure: The impact of changing BF machine on the mean AIQ scores of agents.

The Influence of MC-AIXI Parameters on its AIQ

- *720 configurations (decay (D)=true) with:*
  - number of Monte Carlo simulations (MC): 50, and 100,
  - context tree depth (CTD): 8, 16, and 32,
  - agent horizon (AH): 1, 2, 3, 4, 5,
  - exploration (E): 0.8, 0.85, 0.9, 0.95,
  - exploration decay (ED): 0.3, 0.6, 0.9, 0.95, 0.99, 0.995.
- *120 configurations (decay (D)=false) with:*
  - MC, CTD, AH as above,
  - exploration (E): 0.05, 0.1, 0.15, 0.2,
  - exploration decay (ED): 1 (no decay).
- **Using the default settings experiment**.
- *The analysis of MC-AIXI results:*
  - Box plots showed differences among parameter-values group means.
  - **Significant in all parameter groups** (repeated measures ANOVA).
  - Initial **Classification and Regression Trees** (CART) and **Conditional Inference Trees** showed more nuanced influences.

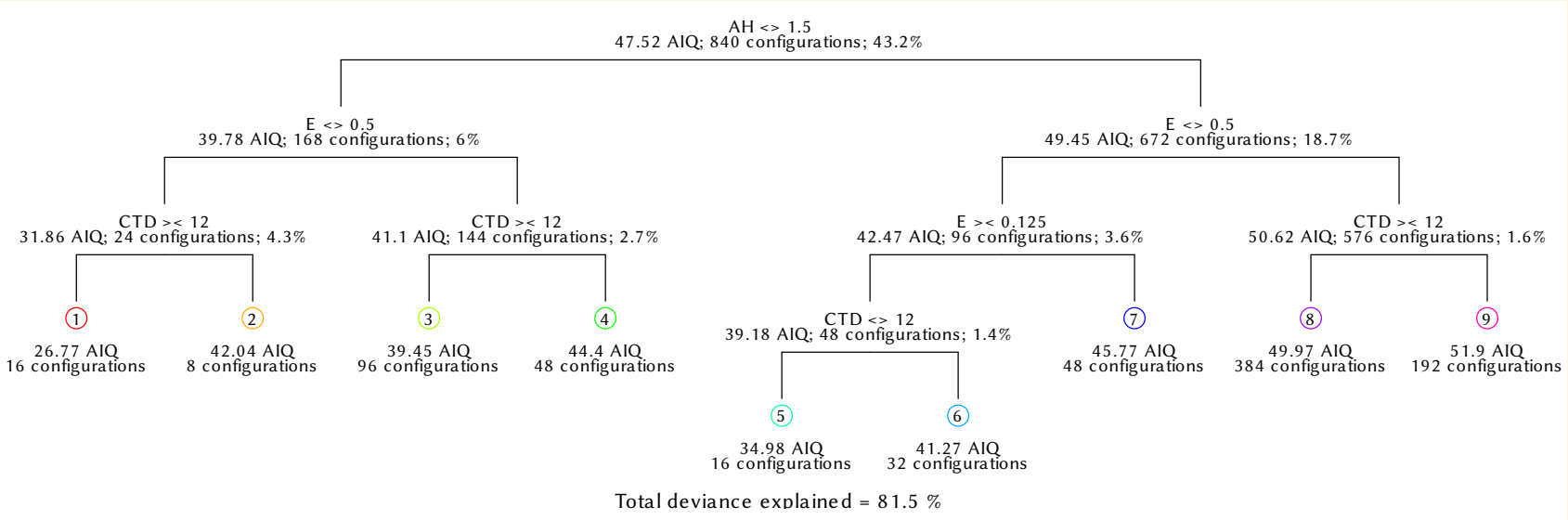

Figure: Regression tree showing the impact of changing MC-AIXI parameters on its AIQ for EL of 100,000 interactions. (Interpret node `E <> 0.5` as `D False × True`.)

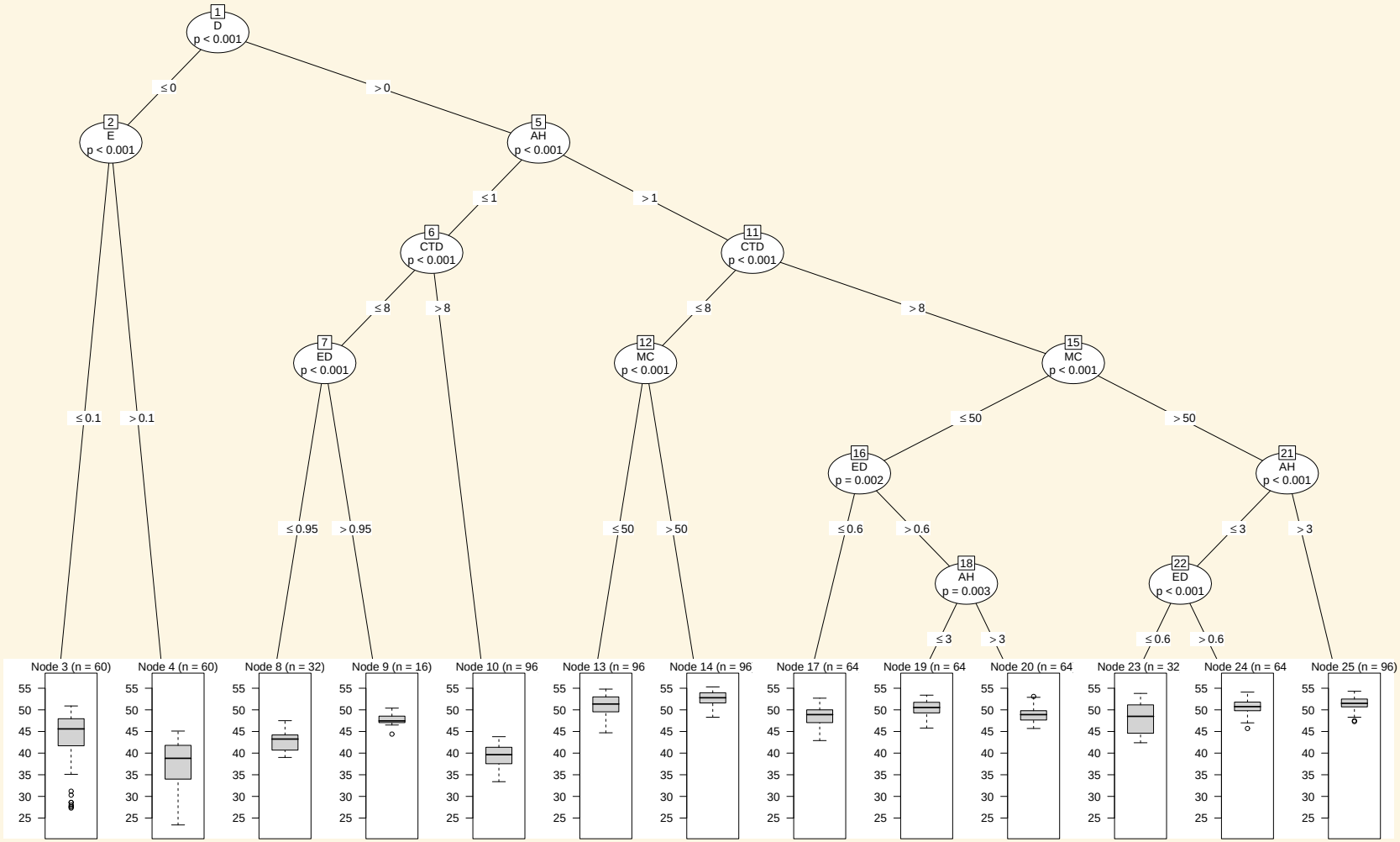

Figure: Conditional inference tree showing the impact of changing MC-AIXI parameters on its AIQ for EL of 100,000 interactions.

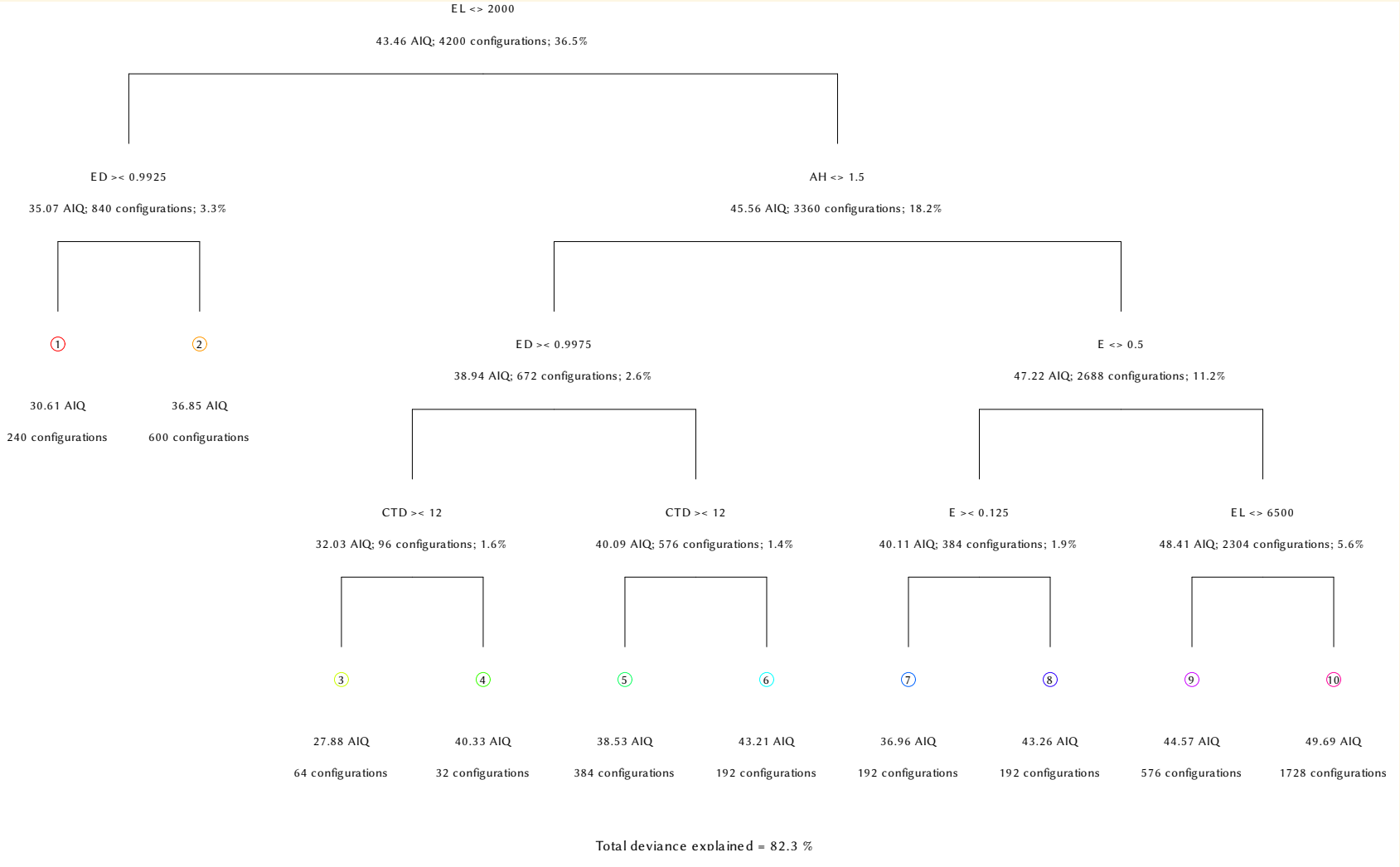

Figure: Regression tree showing the impact of changing MC-AIXI parameters on its AIQ for all ELs. (Interpret node `E <> 0.5` as `D False × True`.)

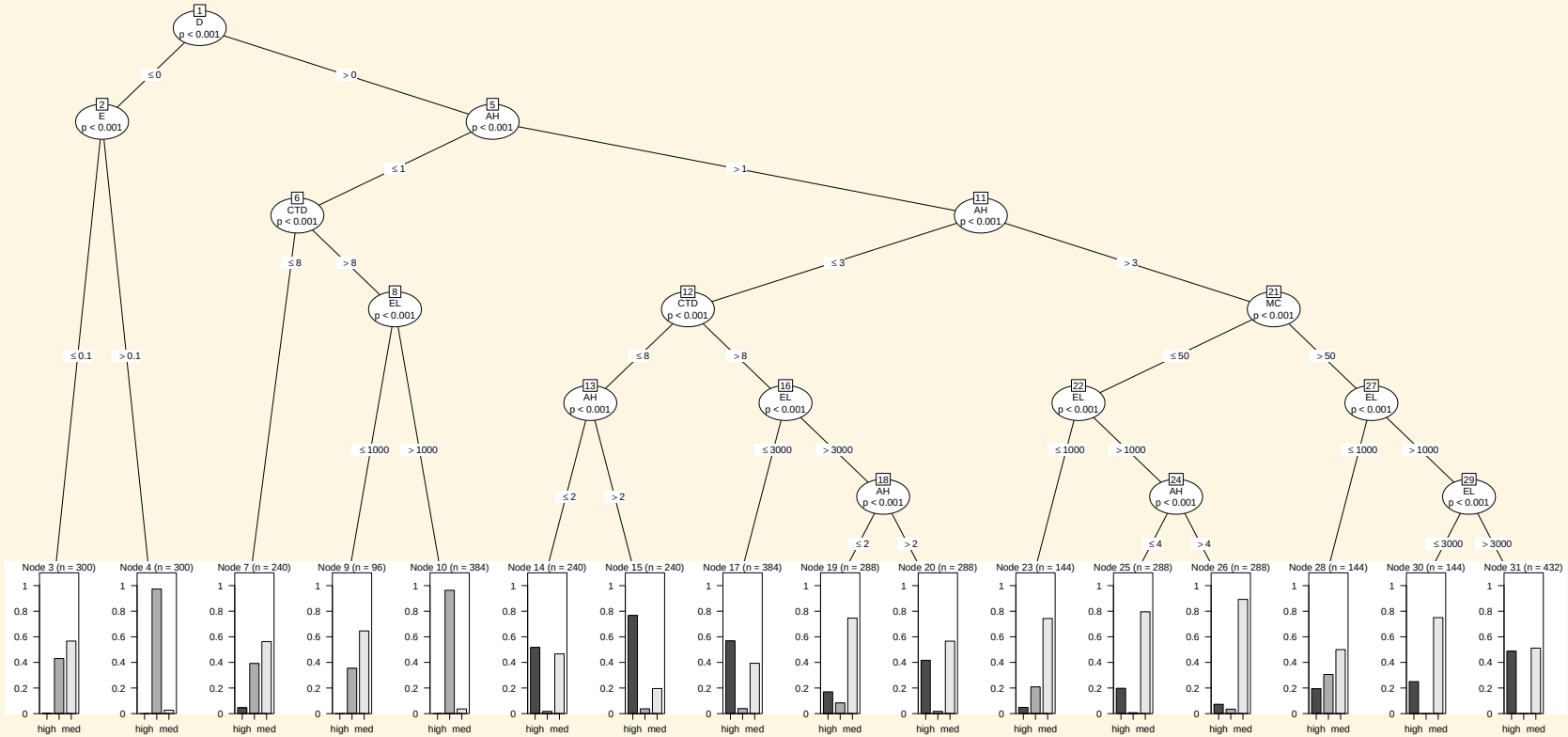

Figure: Conditional inference tree showing the impact of changing MC-AIXI parameters on its AIQ for all ELs. AIQ classified as high (above 3rd quartile for given EL), low (below 1st quartile for given EL), and medium (between 1st and 3rd quartiles for given EL).

AIQ Test Insights

- **The test is computationally rather demanding**:
  - number of tested environment programs,
  - sizes of action and observation spaces,
  - episode length.
- **Large configuration space of agents**:
  - number of agent parameters,
  - number and type of parameter values.
- **Pronounced differences among agent configurations**:
  - statistical reporting of results.
- **Lesser differences among the overall results of agents**.
- **Interpretation of AIQ test results**:
  - theoretical × practically achievable maximal AIQ,
  - classes of environment programs,
  - the actual role of observations.

Semantic Analysis of Environment Programs

- **Understanding environment programs**:
  - How does chance/agent's action influence rewards/observations?
  - Is observation information provided?
  - What is the influence of simple programs?
  - Is there pointless code in environment programs?
  - ...
- **Identify classes of environment programs**:
  - reward is surely based on chance,
  - no observation is provided,
  - only read and write instruction is used,
  - action is overwritten by chance.
  - ...
- **Current methods** (better methods?):
  - identify possible syntax for given semantics,
  - describe by a regular expression.
- **Impact of environment programs classes on AIQ of agents?**

Extending the AIQ Test

- Implementing concepts from Goertzel's critique:
  - **computational efficiency**,
  - **intellectual breadth**.
- **Dynamics of AIQ score convergence**.
- **Technical improvements**:
  - ineffectiveness of computing results on several EL,
  - ineffectiveness of computing more precise results with higher number of environment programs,
  - environment programs ineffectiveness.

Employing the AIQ Test

- **Methodology to evaluate intelligence of artificial systems**:
  - combining Universal Intelligence definition for formal analysis,
  - AIQ test for practical testing,
  - and allowing for Goertzel's critique:
    - establishing computational efficiency,
    - and degree of intellectual breadth
- **Evaluation of AI systems and paradigms**:
  - deductive approach from a formal analysis using Universal Intelligence definition,
  - inductive approach from results of actual AIQ tests,
  - generalization from AI systems to AI paradigms,
  - answering questions regarding foundational concepts of AI.

Conclusion

- *Overview of approaches defining, and evaluating intelligence:*
  - Turing test extensions, and intelligence interconnected with cognitive abilities,
  - **Universal Intelligence definition** and **AIQ test**,
  - **computational efficiency** of intelligence and **intellectual breadth** of an agent.
- *Replicating experiments with AIQ test:*
  - default and varying the observation space matches the original, however varying the action space seems to have some impact,
  - differences in MC-AIXI results due to the exact original parameters being not known,
  - analysis of the influence of MC-AIXI parameters on its AIQ.

Future Work

- **More test of AI systems:**
  - more thorough tests of tested agents to determine parameters influence,
  - other agents for better comparison.
- **AIQ test improvements:**
  - technical improvements,
  - incorporating Goertzel's critique,
  - results of environment programs analysis.
- **Semantic analysis of environment programs:**
  - classes of environment programs and their impact,
  - factors influencing the rewards,
  - the role of observations,
  - pointless and ineffective code.
- **Methodology to evaluate intelligence of artificial systems.**

Acknowledgment

Computational resources for the experiments were kindly provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the program "Projects of Large Research, Development, and Innovations Infrastructures".