Journal of Artificial General Intelligence 9(1) 1-54, 2018
DOI: 10.2478/jagi-2018-0001

Submitted 2017-02-17
Accepted 2018-02-06

Towards General Evaluation of Intelligent Systems: Lessons Learned from Reproducing AIQ Test Results

Ondřej Vadinský ONDREJ.VADINSKY@VSE.CZ
Department of Information and Knowledge Engineering
University of Economics, Prague, Czech Republic

Editor: José Hernández-Orallo

Abstract

This paper attempts to replicate the results of evaluating several artificial agents using the Algorithmic Intelligence Quotient test originally reported by Legg and Veness. Three experiments were conducted: One using default settings, one in which the action space was varied and one in which the observation space was varied. While the performance of *freq*, Q_0, Q_λ, and HLQ_λ corresponded well with the original results, the resulting values differed, when using *MC-AIXI*. Varying the observation space seems to have no qualitative impact on the results as reported, while (contrary to the original results) varying the action space seems to have some impact. An analysis of the impact of modifying parameters of *MC-AIXI* on its performance in the default settings was carried out with the help of data mining techniques used to identifying highly performing configurations. Overall, the Algorithmic Intelligence Quotient test seems to be reliable, however as a general artificial intelligence evaluation method it has several limits. The test is dependent on the chosen reference machine and also sensitive to changes to its settings. It brings out some differences among agents, however, since they are limited in size, the test setting may not yet be sufficiently complex. A demanding parameter sweep is needed to thoroughly evaluate configurable agents that, together with the test format, further highlights computational requirements of an agent. These and other issues are discussed in the paper along with proposals suggesting how to alleviate them. An implementation of some of the proposals is also demonstrated.

Keywords: artificial general intelligence, evaluating intelligence of artificial systems, Universal Intelligence definition, Algorithmic Intelligence Quotient test

1. Introduction

The recently established field of artificial general intelligence (AGI), also referred to as strong artificial intelligence, focuses on understanding and developing artificial intelligence comparable to that of a human, especially with respect to its universality, cf. Searle (1980); Goertzel (2014). This is in agreement with the original question of artificial intelligence (AI): "What is intelligence?" noted already by Turing (1950). As the field has evolved, this question has become increasingly ignored. Recently though, it came back into focus in attempts to define and later also test and measure intelligence made by Legg and Hutter (2007b); Hernández-Orallo and Dowe (2010); Insa-Cabrera et al. (2011); Legg and Veness (2013). As a result of this work, a new research area has emerged, focusing on universal evaluation of intelligence. Its overview is given by Hernández-Orallo (2017).

Through philosophical reflection and the knowledge of cognitive science, the notion of intelligence can be grasped and framed onto other high-level concepts describing related abilities of the mind. However, such an approach provides neither a well-specified definition of intelligence, nor

a practical method by which to evaluate it. There are some methods of practical intelligence testing in psychology that can be adopted by Psychometric AI to evaluate artificial systems (Bringsjord and Schimanski, 2003). A lot of work has been done recently to try and achieve a well-specified definition of intelligence using Algorithmic Information Theory. Most notably, it has resulted in the Universal Intelligence definition (Legg and Hutter, 2007b). Finally, there are approaches that combine the features of being both well specified and practical in terms of evaluation procedure: the Anytime Intelligence test (Hernández-Orallo and Dowe, 2010), and the Algorithmic Intelligence Quotient test (Legg and Veness, 2013). More details will follow in Section 2.

The Algorithmic Intelligence Quotient test (AIQ test) enables practical evaluation of intelligent agents (Legg and Veness, 2013). The original paper focused on the way in which the test was derived from the Universal Intelligence definition (Legg and Hutter, 2007b). Results achieved by several agents were also presented. This paper will first focus on reproducing those results (in Section 3), and then assess the AIQ test regarding its suitability as a general AI evaluation method in Section 4. Based on the conducted experiments and overviewed literature, existing limits of the test will be identified and possible methods to alleviate them will be proposed and discussed.

2. Defining and Measuring Artificial General Intelligence

Philosophical and cognitive presumptions of intelligence will be summarized in Section 2.1. The ideas of Psychometric AI will be outlined in Section 2.2. The Universal Intelligence definition will be introduced in Section 2.3, while the Pragmatic General Intelligence definition will be briefed on in Section 2.4. The Anytime Intelligence test will be introduced in Section 2.5. Finally, the Algorithmic Intelligence Quotient test will be detailed in Section 2.6.

2.1 Philosophical and Cognitive Presumptions of Intelligence

Philosophical reflections on the meaning of intelligence and thought can be traced back to at least Descartes (1637), who points out its universality and its connection to language and rational speech. Searle (1980) notices the relationship between intelligence, meaning, understanding and intentionality. However, intelligence can be delimited even more widely, e.g. in relation to consciousness, as discussed in Dennett (1991). As, for example, de Mey (1992) points out, intelligence requires knowledge (a representation of the world in some kind of model) which is formed as the system interacts with the world through perception and action. The connection of intelligence to cognitive abilities in humans has been extensively modeled by cognitive architectures, as overviewed e.g. by Sun (2007).

Based on such framing of intelligence in high-level concepts related to other cognitive abilities, the initial work in the area of AI evaluation can be said to have begun with the well-known Turing test (Turing, 1950). Arguably, it explicitly tests some of the language capacities of the evaluated entity, while implicitly it considers evaluation of other capacities which can be reported by language. Similar, expanded tests of intelligence were later proposed e.g. by Harnad (1991), who focuses on the full repertoire of human intelligent behavior, and Schweizer (2012), who accentuates the ability of a species to evolve its intelligent behavior. These tests have many problems, however, their evolution shows a shift from unspecified AI to agents interacting with environments (and possibly even other agents), and a tendency to evaluate them against a growing and more explicit set of tasks. This trend is also advocated by Hernandez-Orallo (2000), who gives the motivation to switch from the Turing test to tests and definitions based on the Algorithmic Information Theory.

2.2 Psychometric Artificial Intelligence

Psychometric AI (PAI) of Bringsjord and Schimanski (2003) approaches Turing's (1950) question: "What is intelligence?" from an angle that stresses the testing aspect.

Psychometrics is a field of psychology that deals with the systematic measurement of psychological properties (especially intelligence) in humans using various tests. According to Bringsjord and Schimanski (2003) it gives an answer to the question of what intelligence is, and therefore all AI research should be understood as PAI and should focus on "building information-processing entities capable of at least solid performance on all established, validated tests of intelligence and mental ability." This results in an iterative approach in which the ability to solve an increasing number of tests is integrated into the entity in question – which is useful from an engineering perspective.

However, as a testing approach, PAI is somewhat impractical, since it deals with 'all established and validated tests', which is an open set. It also does not explicitly address the issue of defining intelligence, leaving that to psychology. That may be considered a benefit by some, however, questions can be raised as to whether psychological definitions of human intelligence implicitly present in the tests can be directly applied to an artificial system. A more general approach supported by multidisciplinary interaction may be needed. The limitations of PAI are also considered by Besold, Hernández-Orallo, and Schmid (2015) who state several arguments questioning the necessity and sufficiency conditions of directly using human intelligence tests to measure machine intelligence. They also call for generalization and improvement of tests used by PAI.

2.3 Universal Intelligence Definition

In order to answer the question: "What is intelligence?" with a precise definition, Legg and Hutter (2007a) studied a broad variety of definitions, theories, and tests of human, animal, and artificial intelligence and arrived at the following (informal) result: "Intelligence measures an agent's ability to achieve goals in a wide range of environments." Legg and Hutter (2007b) give its formalization as shown by Equation 1.

$$\Upsilon(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_\mu^\pi \tag{1}$$

where the Universal Intelligence Υ of agent π is given by its ability to achieve goals as defined by a value function $V_\mu^\pi := \mathrm{E}\left(\sum_{i=1}^{\infty} r_i\right) \leq 1$ basically as maximizing the expected sum of all future rewards (given a history of interactions) over a set E of environments μ weighted by algorithmic probability that uses Kolmogorov complexity K. See details in Legg and Hutter (2007b).

Looking at the formal definition by Legg and Hutter (2007b) in Equation 1, the following important building blocks can be noted:

- The definition considers the environment μ, agent π, and their iterated interaction through actions a_i of the agent, and its perceptions (observations) o_i and rewards r_i originating in the environment. The environment is defined as a Turing computable conditional probability measure μ of perceptions and rewards given current interaction history.

- With all computable environments considered, Kolmogorov complexity $K(\mu)$ is used as a measure in place of Occam's razor $2^{-K(\mu)}$. If a short program can be used to describe the probability measure of an environment, than the environment has a low complexity, since Kolmogorov complexity is based on the length of the shortest program describing a sequence

of bits. Thus, complex environments are less influential on the agent's overall performance than simple ones.

- An agent's ability to achieve goals is described by a value function V_μ^π as maximizing the expected future rewards given past interaction with the environment. Temporal preference is present in the way rewards are distributed by the environment. That is, the environment itself decides if a slow-learning but more accurate or fast-learning but inaccurate solution is better.

The Universal Intelligence definition of Legg and Hutter (2007b) can be seen as a generalization of the earlier work of Hernandez-Orallo (2000) on C-tests from static to dynamic environments.

The definition of Legg and Hutter (2007b) orders the performance of agents, ranging from unintelligent random behaviour to theoretically optimal AIXI. When considering all computable environments, weighted by their complexity, only a truly general agent can attain high Universal Intelligence. Founded on concepts of information theory, it is not culturally biased or anthropocentric. As could be noted from previous paragraphs, this definition is not computable due to it considering infinitely many environments, infinitely long agent – environment interaction, and uncomputable Kolmogorov complexity. As a result, after dealing with those three limitations, any measure based on this definition can only be approximate.

Although culturally unbiased, the Universal Intelligence of an agent is, through the use of Occam's razor $2^{-K(\mu)}$, dominated by a fairly small set of short programs describing simple environments. Program length is assessed by the Kolmogorov complexity function K relative to a reference Turing machine $\mathcal{U}$. The choice of the reference machine (i.e. the programming language) determines which classes of environments (or problems) can be described by short programs. Therefore, Universal Intelligence can be made biased towards a certain class of environments by the choice of a suitable reference machine. Legg and Hutter (2007b) consider this to be an issue especially for simple agents, however Hibbard (2009) showed that under specific conspirative conditions it can cause serious issues even for the optimal agent AIXI. According to Hibbard, this bias can be arbitrarily reduced by specifying the minimal length of environment programs considered by the definition. Hernández-Orallo (2015, 2017) proposed an alternative solutional approach based on the idea that it is in fact the complexity of the solution that determines the difficulty of the problem (environment). He therefore suggests changing the way overall score is aggregated so that it incorporates this idea of environment difficulty, defined as Levin's Kt complexity of the simplest solution for the problem.

2.4 Pragmatic General Intelligence Definition

A critique of Universal Intelligence Definition stressing its applicability on real agents in real environments was given by Goertzel (2010). The critique is based on the following three aspects:

- There are implicit goals, and goals set by the agent. Also, rewards do not necessarily come from the environment, and not all intelligent behavior is goal and reward oriented. To mitigate this issue, Goertzel's definition of Pragmatic General Intelligence explicitly considers goals.

- Agents are usually adapted to particular environments to some degree, therefore a somewhat biased generality can be of interest (especially for comparison with humans). Thus, Goertzel's definition allows for other environment probability distributions than Universal distribution.

- Real-world agents have to operate with limited resources, therefore efficiency of intelligence is of practical importance. Consequently, Goertzel introduces the Efficient Pragmatic General Intelligence definition which explicitly considers consumption of computational resources.

Related to the definitions above is also the initial attempt of Goertzel (2010) to formally specify what general versus specific means in terms of intelligence. He calls it the Intellectual Breadth of an agent and proposes to use a fuzzy set of contexts, i.e. environments, goals, and time intervals, relative to which the agent is intelligent. This resulting fuzzy set can be normalized to a probability distribution and some measure, e.g. entropy, can then be used to assess it.

2.5 Anytime Intelligence Test

The Anytime Intelligence Test (AIT) is an intelligence test proposal for present and future artificial and biological agents of any intelligence level working at any time scale introduced by Hernández-Orallo and Dowe (2010). The test can be stopped at any time, producing results whose accuracy improves with the duration of the evaluation. The test is based on the Universal Intelligence definition by Legg and Hutter (2007b), modified so that it is computable, and then combined with the earlier work of Hernandez-Orallo (2000) on C-tests, and that of Dowe and Hájek (1998) on an induction enhanced Turing test.

Hernández-Orallo and Dowe (2010) deal with the three dimensions of uncomputability of the Universal Intelligence definition in the following ways:

- A sample of environments is used instead of all environments, raising the question of their discriminative power. Unlike in the original definition, Hernández-Orallo and Dowe propose that only reward-sensitive environments be included, thus excluding environments that might ignore an agent's behaviour entirely. Therefore, the test uses only the environments in which an agent's behaviour can always have an impact on its rewards.

- A limited number of agent – environment interactions is used instead of infinitely many. This, however, raises a question of how to suitably combine rewards into a single score. Hernández-Orallo and Dowe propose averaging rewards by the number of interactions. Also, they suggest using balanced environments with rewards ranging from -1 to $+1$ which would cause a randomly behaving agent to score (on average) zero.

- A bounded and computable version of Kolmogorov complexity originating in Levin's Kt is used as a distribution function for the environments. The function is called $Kt^{\max}$ and enforces a time limit on the environment for the computation of its interaction with an agent.

Moreover, Hernández-Orallo and Dowe (2010) addressed two other aspects they consider important in an intelligence test: adaptiveness of the testing process, and the relationship between time and intelligence. Physical time is incorporated into the test by setting a time limit for agent – environment interaction, as well as by integrating the time limit into the computation of the overall score. The testing process of the Anytime Intelligence Test is adaptive in the sense that the agent is tested on environments of progressively increasing or decreasing complexity, as well as with a progressively increasing or decreasing time limit, in order to effectively match an agent's intelligence level as well as time scale. The reference machine used by the test is explicitly stated as its parameter and it can range from a very restricted state automata to universal Turing-complete machines.

Insa-Cabrera et al. (2011) introduced a prototype implementation of a simplified version of the AIT by Hernández-Orallo and Dowe (2010) using an (also simplified) version of the Unbiased Universal Environment class by Hernández-Orallo (2010) as a reference machine and conducted an experiment with it comparing humans to a Q-learning based artificial agent. This was possible due to the idea of having different subject-specific interfaces in front of the same test. Interfaces change the representation of rewards, actions, and observations based on the type of subject.

This prototype implementation of Anytime Intelligence Test by Insa-Cabrera et al. (2011) uses simple state-space based environments of varying size which the agent can navigate through via its actions. Rewards and penalties are generated by two processes (agents in the original terminology) which navigate the environment using a fixed pattern of actions in a loop and a random starting point, ensuring the environments are balanced and reward sensitive. The length of an LZ compressed string of actions used to generate rewards and penalties is used as a complexity estimate of the environment, making the estimate feasible, however also rather distant from $Kt^{\max}$ of the AIT proposal or Kolmogorov complexity of the Universal Intelligence. An agent is allowed a limited number of interactions based on the number of states of the environment. However, the time scale of the agent is not considered by the prototype, missing one of the key aspects of the AIT proposal. Also, the environments are not generated by a Turing-complete process, and are wholly observable by the agent, therefore only a subset of the environments considered by the Universal Intelligence definition can be used by the prototype implementation of AIT.

2.6 Algorithmic Intelligence Quotient Test

The Algorithmic Intelligence Quotient, proposed by Legg and Veness (2013) and shown in Equation 2, is a computable approximation of Universal Intelligence that can be be tested practically. It is defined as:

$$\hat{\Upsilon}(\pi) := \frac{1}{N} \sum_{i=1}^{N} \hat{V}_{p_i}^{\pi} \tag{2}$$

where the AIQ $\hat{\Upsilon}$ of agent π is given by its ability to achieve goals as defined by an empirical value function $\hat{V}_{p_i}^{\pi}$ as a total reward from a single trial of an environment program p_i averaged over N sampled programs. See details in Legg and Veness (2013).

Looking at Equation 2 (Legg and Veness, 2013), the following ways of dealing with the three dimensions of the uncomputability of the Universal Intelligence definition can be noted:

- The test considers a finite sample of N environment programs p_i, agent π and their interaction. The same environment can be described by several programs and the same program can be included in the sample many times in order to account for its higher weight.

- With N environment programs considered, a simple average is taken. However, the notion of Occam's razor is kept in the way environment programs are sampled since Solomonoff's Universal Distribution is used: $M_{\mathcal{U}}(x) := \sum_{p:\mathcal{U}(p)=x*} 2^{-l(p)}$. Therefore, a shorter program has a higher probability of being selected, but all programs describing an environment are considered, not only the shortest (as is the case with Kolmogorov complexity). This switch of distribution is closer to the original definition than the solution of Insa-Cabrera et al. (2011).

- An empirical value function $\hat{V}_{p_i}^{\pi}$ is used since only a limited number of iterations are tried. The rewards given by the environment program are no longer bound by 1 as is the case with

the definition, nor are they in any way discounted to specify the temporal preference. That is, the total reward returned from a single trial of agent – environment interaction is used.

An environment program is, unlike the environments used by Insa-Cabrera et al. (2011), a Turing-complete program that computes the current reward and observation from the interaction sequence. As is the case with the Universal Intelligence definition, results of the AIQ test depend on the choice of a reference machine, since it influences which classes of environments are more likely to be sampled. In an attempt to minimize this dependency, the test uses a rather simple BF reference machine by Müller (1993) that has been extended so that it can also write a random symbol, thus allowing for indeterminism. Adopting the call for balanced environments by Hernández-Orallo and Dowe (2010), computed rewards are normalized to the interval $[-100, +100]$ fixing minimal and maximal AIQ while ensuring a randomly behaving agent will score 0. Action and observation symbols as well as internal states of environments are moduloed integers. The machine uses a one-way read-only input tape for an agent's current action with a history of 24 previous actions, a two-way read-and-write work tape of 100,000 cells in each direction, and a one-way write-only output tape for a reward and a configurable number of observations. Due to this design, environments that are not fully observable are likely. The BF language uses 10 instructions:

- `+-` increment/decrement respectively the symbol on the working tape,
- `,.` read from an input tape and write to the current cell of the work tape/write the current cell of the work tape to the output tape respectively, and move the respective input or output tape pointer to the right,
- `<>` move the work tape pointer to the left or right,
- `[]` start a loop if the current work cell is non-zero/end the loop respectively,
- `%` write a random symbol to the current work cell,
- `#` end program.

To account for non-halting and long-running programs, the computation of each iteration is limited to 1,000 steps. This is further encouraged by halting the program if it tries to write more than the set number of reward and observation symbols. Aside from excluding long-running programs, the proportion of non-interactive programs (described as 'passive' in the authors' terminology) is also considerably reduced by mandatory read and write instructions and by omitting programs that return constant rewards (Legg and Veness, 2013, 2011). Because of that, the call of Hernández-Orallo and Dowe (2010) for excluding non-discriminative environments is partially satisfied.

Legg and Veness (2013, 2011) use several techniques to reduce the variance and to speed up the AIQ estimation process. One of them is adaptive stratified sampling, which classifies environment programs into 20 mutually exclusive strata. 10 of these are based on the presence of simple patterns in returned rewards, and the other 10 are classified by program length. An agent is tested on programs that are chosen in order to maximally minimize the variance of the agent's AIQ estimate in a given stratum, resulting in sampling more programs from strata on which the agent has more varied results.

An open source prototype implementation of the AIQ test is available from Legg and Veness (2011). In the test, the *size of the environment programs sample* used by the adaptive stratified

estimator can be configured, impacting the precision of AIQ score estimates. The *number of agent – environment interactions* influences an agent's learning time for the trial and if set sufficiently high enables its score to converge. Moreover, *sizes of observation and action space* can be configured, affecting the interaction space complexity. Also, the *number of returned observation symbols* can be set, further increasing the complexity of the interaction space as well as possibly prolonging the environment computation by increasing the write limit (Legg and Veness, 2013, 2011).

The test can be performed on agents supplied internally or externally via a custom wrapper. The following agents are provided with the test: *random*, *freq*, Q_λ by Watkins (1989) subsuming Q_0, HLQ_λ by Hutter and Legg (2007), and a wrapper for a Monte Carlo AIXI approximation (*MC-AIXI*) by Veness et al. (2011). The agent *random* performs random actions. The agent *freq* has a parameter ϵ for an ϵ greedy action selection, otherwise it chooses the action with the highest average reward. The agent Q_λ has several parameters: *initial* Q, eligibility trace discounting rate λ, learning rate α, ϵ, and a discounting rate γ. The agent HLQ_λ has the following parameters: *selection mode* to enable (0) or disable ϵ greedy selection, *initial* Q, λ, ϵ, and γ. *MC-AIXI* itself has many parameters, however the wrapper only allows for the configuration of: *number of Monte Carlo simulations* (affecting prediction power), *context tree depth* (affecting the size of an agent's model), *search horizon* (impacting the expectation look-ahead), and *exploration* (effectively being ϵ) (Legg and Veness, 2013, 2011). In the experiments performed by Legg and Veness, however, a further parameter *exploration decay* was used, as clarified by personal inquiry. This allowed for the optional enabling of an exponential decay of ϵ. The *MC-AIXI* wrapper was extended to enable configuration of *exploration decay*.

3. Reproducing the Results of AIQ Test

In order to assess the AIQ Test, an experiment will be conducted in Section 3.1 in accordance with the default settings reported by Legg and Veness (2013). Further settings mentioned by the paper will also be recreated: varying the size of the action space available for agents in Section 3.2, as well as varying the size of the observation space supplied to agents in Section 3.3.

3.1 The Default AIQ Test

The default AIQ Test uses a BF reference machine with 5 symbol action and observation space, and returns 1 reward symbol and 1 observation each iteration (BF 5). The original results were provided by Legg and Veness (2013) as a plot of the best values achieved by each agent. More details were given by Legg and Veness (2011) including the exact values achieved by tested agents, the configurations of the agents, and samples of 200,000 environment programs used in the experiment.

3.1.1 HYPOTHESES

Given the availability of the original results, and the way they were presented, the following hypotheses were formulated:

- Weak interpretation – *The ordering of agents according to their maximal AIQ is the same as that given by Legg and Veness (2013).*
- Strong interpretation – *The maximum AIQ scores of the agents are not significantly different from that given by Legg and Veness (2011).*

Table 1: Statistics of sampled environment programs for reference machines used in all experiments including the original samples (BF 5_{LV}) of Legg and Veness (2013).

Measure	**BF 2**	**BF 5**	**BF 5_{LV}**	**BF 10**	**BF 20**	**BF 52**	**BF 53**	**BF 54**
Mean	20.62	21.07	21.18	20.90	20.77	21.02	21.01	20.89
Standard Error	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Minimum	3	3	3	3	3	3	3	3
First Quartile	7	8	8	7	7	8	8	8
Median	13	14	14	14	13	14	14	14
Third Quartile	25	26	26	26	25	26	26	26
Maximum	476	362	355	424	523	369	479	399
Standard Deviation	22.50	22.25	22.51	22.28	22.20	22.18	22.07	21.78
Kurtosis	21.08	17.01	17.01	17.53	19.52	18.03	20.00	18.93
Skewness	3.47	3.23	3.24	3.29	3.37	3.30	3.37	3.31

3.1.2 SETTINGS

For the experiment, 200,000 new environment programs were generated using the sampler provided by Legg and Veness (2013). The new sample file includes 161,205 unique programs, similar to the 161,069 unique programs in the original sample file. 6,554 unique programs are shared by both samples. The resulting program statistics is given in Table 1 according to program length.

In the sample there are many simple programs like `[+.>]-,,#` which writes the incremented previous action symbol as a reward if it was non zero. Some programs ignore agent's actions like `,%.#` which just writes a random reward. The same environment is described by `,,%.+,%#` which does some more meaningless operations. There are also more complex programs like:

```
+[[[+>]<<.,<%->]%,>-<],+[+>.,.%]+>.%%....<+#
```

With longer programs however, some of the complexity can be lost, as the execution ends if more than a reward and a set number of observations is written – i.e. in the example above, the program actually ends with third executed dot (when using BF 5 reference machine).

Since each agent has several parameters, it would be optimal to do a full sweep through the configuration space of each agent for all tested episode lengths. This did not prove feasible since the configuration space of agents is large and a single test run is computationally demanding. Therefore, the same settings were used as Legg and Veness (2011) reported for *freq*, Q_0, Q_λ, and HLQ_λ. For *freq* two additional configurations (*#1* and *#5*) were devised so as to have the same number of tested configurations for each agent. A full list of configurations of these agents is given in Table 7.

For *MC-AIXI*, however, there was only a single configuration given by Legg and Veness (2011) with 100 *Monte Carlo simulations*, *context tree depth* of 32, and a *search horizon* of 3. With nothing else specified, this resulted in a default ϵ value of 0.05 being set by the wrapper. This, however, was not the configuration used by Legg and Veness in their experiments, as clarified by a personal inquiry. They varied the ϵ values, and also used an additional *exploration decay* parameter, however they could no longer provide exact values, only point out certain regions. Based on this information and inspired by an agent-scaling experiment of Legg and Veness (2013), a parameter sweep for *MC-AIXI* was attempted by testing all combinations of the following values:

- *number of Monte Carlo simulations*: 50, and 100,
- *context tree depth*: 8, 16, and 32,
- *search horizon*: 1, 2, 3, 4, and 5,
- *exploration*: 0.8, 0.85, 0.9, and 0.95,
- *exploration decay*: 0.3, 0.6, 0.9, 0.95, 0.99, and 0.995.

This way, high exploration values with both slower and faster decay rates were tested. Also, configurations with low *exploration* values 0.05, 0.1, 0.15, and 0.2 were tried with no decay.

To estimate the AIQ score of an agent, 10,000 environment programs are used for a reasonably small 0.95 confidence interval. As *MC-AIXI* is rather demanding and many configurations were tested, 1,000 programs were used leading to somewhat larger 0.95 confidence intervals. Each configuration was then evaluated after 1,000, 3,000, 10,000, 30,000, and 100,000 interactions with the environment (referred to as *Episode Length*). No whole test runs were discounted, however some agents have parameters influencing their internal discounting as explained in Section 2.6.

3.1.3 RESULTS

Figure 1 shows the best achieved AIQ score estimates with a margin of error corresponding to a 0.95 confidence interval for each agent after the tested number of interactions. To allow for direct comparison with the original results of Legg and Veness (2013), those are also included in the figure. A graphical summary of all *MC-AIXI* results is shown in Figure 2.

Detailed results of tested agents are included in Appendix A. Table 7 gives the full listing of the AIQ score estimates with a margin of error corresponding to a 0.95 confidence interval for each agent configuration after the tested number of interactions. Table 8 lists descriptive statistics of *MC-AIXI* results computed from all tested configurations.

3.1.4 DATA ANALYSIS

As can be seen in Figure 1, the AIQ estimates of *freq*, Q_0, Q_λ, and HLQ_λ correspond well with the results of Legg and Veness (2013). The estimates are within confidence levels in some cases, slightly lower in others, however the ordering of agents remains the same. See Table 2 for a detailed comparison of the differences. Two-sample t statistics (also listed in the table) were computed to determine the significance of the differences. In cases typed in italics, the test rejects that the difference is zero, and the difference is therefore significant.

As for the results of *MC-AIXI*, the situation is more complicated. The best achieved results are substantially better than reported for lower episode lengths (esp. for 1,000 interactions), but considerably worse for higher episode lengths (esp. for 100,000 interactions) – note Figure 1, and Tables 8 and 2. Generally, *MC-AIXI* overtakes Q_λ and catches up with HLQ_λ at higher episode lengths, but never overtakes it as Legg and Veness (2011) reported. While these findings are rather surprising, it should be noted again that, contrary to the analysis of the other agents' results, this one is not based on comparison of the same configurations, since the exact configurations of *MC-AIXI* Legg and Veness used are not known.

Table 3 shows the time needed to test an agent for a given episode length. Only general trends can be compared, since *freq*, Q_0, Q_λ, and HLQ_λ were tested on a quad-core machine (i5-4590

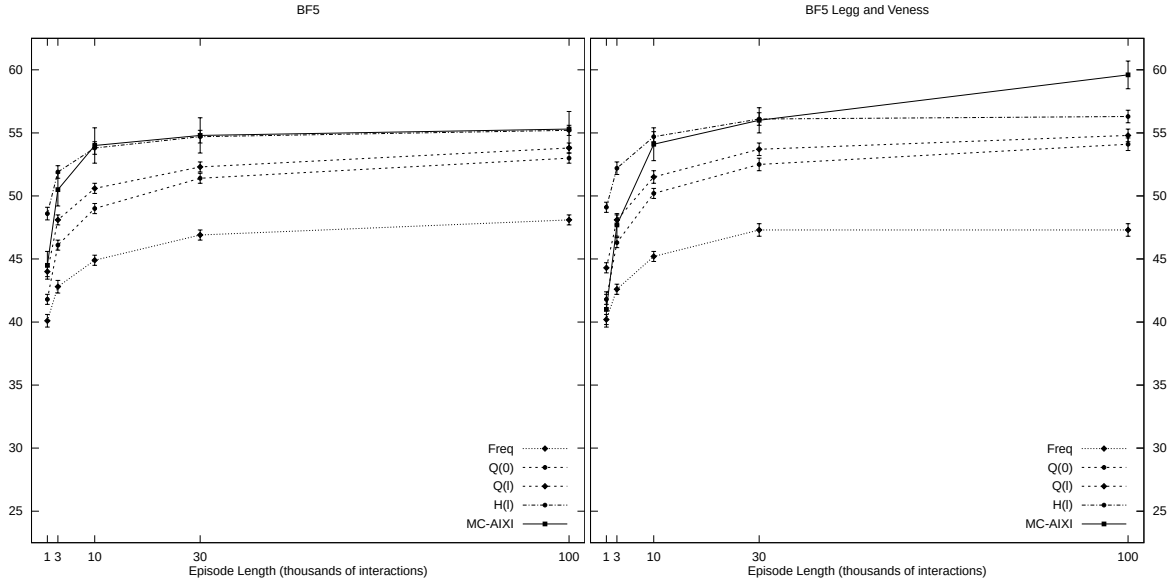

Figure 1: Best achieved estimated AIQ scores of agents as a function of episode length on BF 5 reference machine. To the left are the new results, while the originals of Legg and Veness (2013) are shown to the right.

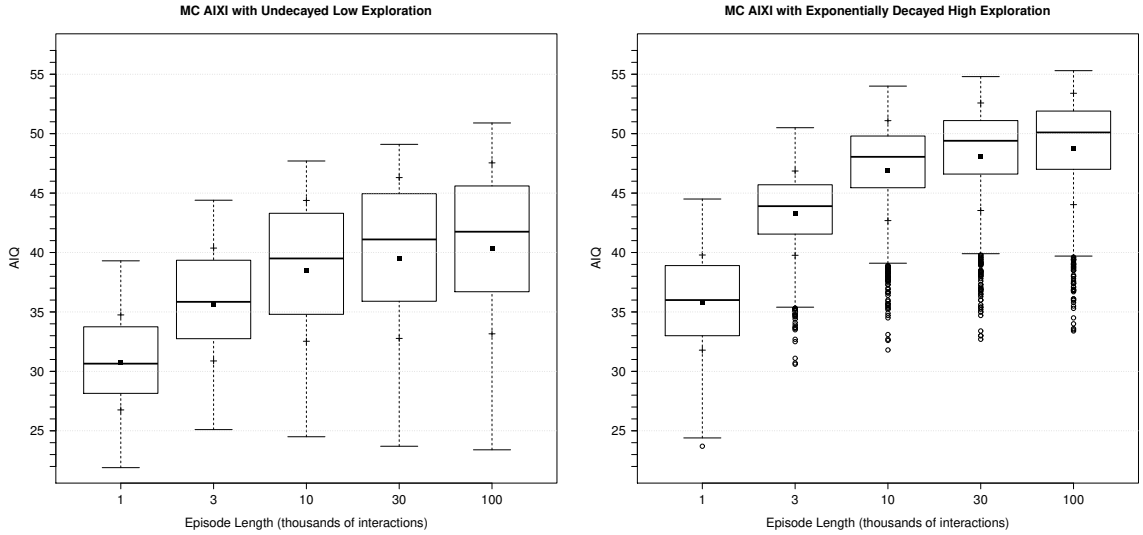

Figure 2: A summary of the estimated AIQ scores of *MC-AIXI* for the configurations without exploration decay (to the left), and for the configurations with exponentially decayed exploration (to the right).

Table 2: Comparison of the best achieved results from Table 7 to the results of Legg and Veness (2013, 2011) – significant differences in italics, extremes in bold (BF 5 Reference Machine).

Agent	**Differences of AIQ Scores with Confidence Intervals for Episode Length**														
	1,000			*3,000*			*10,000*			*30,000*			*100,000*		
	Δ		|t|	Δ		|t|	Δ		|t|	Δ		|t|	Δ		|t|
freq	−0.1	±0.6	0.3	0.2	±0.6	0.6	−0.3	±0.6	1.0	−0.4	±0.7	1.2	*−1.0*	*±0.7*	*3.0*
Q_0	0.0	±0.6	0.0	−0.2	±0.6	0.7	*−1.2*	*±0.6*	*4.0*	*−1.1*	*±0.7*	*3.3*	*−1.1*	*±0.7*	*3.3*
Q_λ	−0.3	±0.6	1.0	0.0	±0.6	0.0	*−0.9*	*±0.7*	*2.6*	*−1.4*	*±0.7*	*4.1*	*−1.0*	*±0.7*	*3.0*
HLQ_λ	−0.5	±0.6	1.6	−0.3	±0.7	0.9	*−0.9*	*±0.6*	*2.8*	*−1.4*	*±0.7*	*4.1*	*−1.1*	*±0.7*	*3.2*
MC-AIXI	***3.5***	***±1.8***	***3.7***	*2.5*	*±1.6*	*3.1*	−0.1	±1.9	0.1	−1.6	±1.7	1.8	***−4.3***	***±1.8***	***4.7***

Table 3: Time required to test AIQ of agents for a given episode length on BF 5 reference machine (*MC-AIXI* not comparable to other agents due to different hardware).

Agent		**Runtime for Episode Length**				
	#	*1,000*	*3,000*	*10,000*	*30,000*	*100,000*
freq	1	00:01:00	00:02:49	00:09:25	00:28:59	01:39:46
Q_0	1	00:01:56	00:06:00	00:19:26	00:57:52	03:16:26
Q_λ	1	00:01:57	00:05:50	00:19:43	00:58:26	03:18:32
HLQ_λ	1	00:05:31	00:17:22	00:57:52	02:46:15	09:53:26
MC-AIXI	1	00:16:15	00:33:12	01:25:12	03:50:27	12:06:38

CPU @ 3.30GHz) with 10,000 samples while for *MC-AIXI* a 16-core node (2× Xeon E5-2650v2 CPU @ 2.60GHz) was used as well as 1,000 samples. While runtimes of *freq*, Q_0, Q_λ, and HLQ_λ are approximately linearly dependent on the tested episode length, runtimes of *MC-AIXI* are noticeably more efficient for shorter episodes (an increase from 1,000 to 3,000 interactions only doubles the runtime) and even for longer episodes maintain a better than linear dependency. This is likely because AIXI builds all consistent models of an environment and then discards those proven inconsistent by further interactions.

3.1.5 SUMMARY OF THE EXPLORATIVE ANALYSIS OF MC-AIXI RESULTS

An analysis was conducted using both statistical and data mining methods to determine the extent to which the results of *MC-AIXI* are influenced by its parameters and their values. While the full analysis can be found in Appendix B, a summary is presented here.

The following general properties of the parameters of *MC-AIXI* in regards to its resulting AIQ were noted by the statistical analysis:

- The overall performance of configurations using *exponentially decaying exploration* (D) is substantially better than that of those using constant exploration.

- Increasing the number of *Monte Carlo simulations* (MC) from 50 to 100 causes only a slight improvement of the results.
- There are also rather limited differences if the *context tree depth* (CTD) is increased from 8 to 16 and to 32.
- Configurations with a *search horizon* (AH) of 1 perform generally rather poorly. When its value is increased to 2 or 3, the results improve considerably, however, further increasing the parameter value has a somewhat negative effect (although this effect diminishes with increasing episode length).
- For configurations not featuring exploration decay, increasing the *exploration* (E) value impedes the performance sizably.
- Configurations with high *exploration decay* (ED) values perform poorly at first, but as the episode length increases, their performance reaches that of the others, and the spread of their results gets much lower than with lower *exploration decay*.

The data-mining analysis confirmed that the parameters *search horizon* and *decay* (a derived parameter denoting whether the exponential decay of exploration is used or not) are the main predictors of the resulting AIQ. However, the parameters' influence is more complex, resulting in the following configurations being identified as especially high performing:

- D $=$ true, MC $= 100$, CTD $= 8$, and AH > 1 at EL $= 100000$;
- D $=$ true, CTD $= 8$, and AH $= 3$;
- D $=$ true, MC $= 100$, CTD $= 32$, AH > 3, E $= 0.8$, and ED $= 0.3$;
- D $=$ true, MC $= 100$, CTD $= 8$, AH $> 1 \leq 3$, and ED ≤ 0.6.

The following configurations were identified as rather poorly performing:

- D $=$ true, CTD > 8, and AH $= 1$ at EL > 1000;
- D $=$ false, CTD $= 8$, and AH $= 1$, at EL $= 100000$;
- D $=$ true, CTD $= 8$, AH $= 1$, and ED ≤ 0.6;
- D $=$ true, MC $= 50$, AH $= 1$, E ≥ 0.85, and ED $= 0.9$.

3.1.6 DISCUSSION

Before hypotheses from Section 3.1.1 can be evaluated, a discussion of the experiment is needed. Since the exact parameters of *MC-AIXI* used by Legg and Veness are not known, a more cautious approach should be taken when evaluating the hypotheses. Instead of considering the results as a whole, the results of *MC-AIXI* will be assessed separately. Based on the data analysis above, and with the mentioned considerations in mind, some hypotheses from Section 3.1.1 were rejected:

- The ordering of agents according to their maximal AIQ *is the same (not considering MC-AIXI)* as that given by Legg and Veness (2013). However, *when MC-AIXI is considered, the ordering differs*.

- The maximal AIQ scores of agents *are in some cases significantly different* from that given by Legg and Veness (2011), however, the differences are rather small for other agents than *MC-AIXI*. In case of *MC-AIXI*, there are more pronounced statistically significant differences.

Keeping in mind the considerations regarding *MC-AIXI*, the differences in its score may well have been caused by the different parameters used in the experiment, and therefore can not form the basis of any overall conclusion. Overall, the replication experiment results correspond well with the findings of Legg and Veness. Therefore, the AIQ test can be said to be reliable, in the sense that in cases when the agent configurations and the test parameters are known the results of testing can be replicated accurately.

As can be seen from Table 7 and the analysis of *MC-AIXI* performance in Section 3.1.5, an agent's parameter values have a great impact on its AIQ, one capable of eliminating the difference between even the best and the worst agent. For HLQ_λ the impact diminishes somewhat at higher episode lengths. The *freq* agent is not greatly affected by alterations to parameter values. Since the tested configurations have rather similar parameters, the differences between the worst and the best can be even greater. Moreover, among the tested configurations, none proved optimal for all episode lengths. These observations stress the need for a thorough parameter search when testing configurable agents. Also, this raises a question of how to evaluate agents with such a vast performance variance.

Comparing the results of agents one to another, there seems to be a rather limited difference. Q_0, Q_λ, and HLQ_λ achieve similar top AIQ scores ranging from 53 to 55. This may not be too surprising, since the agents are all types of Q-learning. The *freq* agent, however, which is of a different and much simpler kind, scores roughly 10–15% lower than the Q-learning agents, with a top score of 48. Also, *MC-AIXI* performs only about 10% better than the Q-learning agents, achieving a top AIQ score of 59.6 (if the results reported by Legg and Veness (2011) are considered). Also, considering the theoretical maximal AIQ of 100 and a minimum of 0 (i. e. demonstrating random behavior), the top results of the agents tested land close to the middle of the range. While it is possible that the actual differences between agents are as minimal as the tests suggests (in which case the test would be valid), it is also possible that the agents actually feature more pronounced differences that the test (at least on the default setting) is too simple to bring out.

As noted during data analysis, test runtimes for *MC-AIXI*, unlike other agents, increased better than linearly with increasing episode length as shown in Figure 3. Similarly, the change in AIQ score convergence rate with increasing episode length differs among tested agents. It is a question of whether these might capture some useful aspect of intelligence, likely something to do with gaining expertise, which should be taken into account when evaluating the agents.

3.2 Varying the Action Space

An experiment manipulating the size of the action space was conducted by Legg and Veness (2013) using 2, 10, and 20 symbol tapes (BF 2, BF 10, and BF 20) with 1 symbol for an observation and 1 for a reward. The results were reported as "qualitatively the same" as when using BF 5, and simply taking longer to run when the number of symbols was increased. No further details provided.

3.2.1 HYPOTHESES

Since a more well-formulated hypothesis is needed than just "results are qualitatively the same", the following specifications were devised:

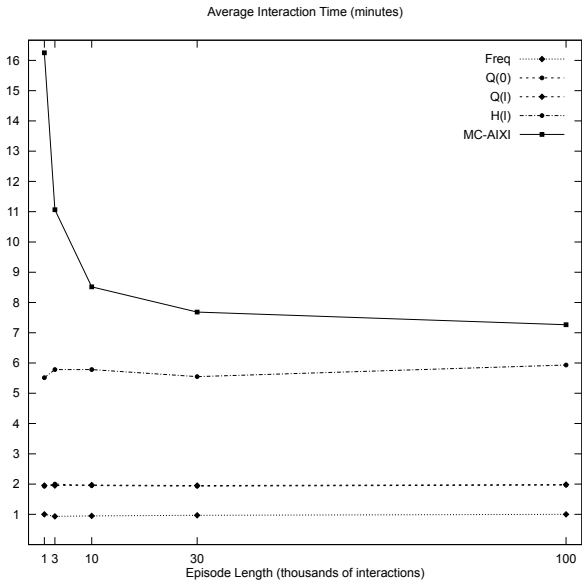

Figure 3: Average time per thousand interactions of *#1* configurations of agents as a function of episode length on BF 5 reference machine (values of *MC-AIXI* are not directly comparable with other agents since different hardware was used).

- Weak interpretation – *The ordering of agents according to their maximal and mean AIQ is the same among the reference machine groups.*
- Strong interpretation – *Group means of AIQ according to a reference machine are not significantly different.*

Since AIQ is a mean over tested environment programs, looking closer at its standard deviation (SD) might provide further insight into the possible differences in the spread of agent performance among tested environment programs when the action space is manipulated. A supporting hypothesis for the strong interpretation can be specified as follows: *Group means of SD of AIQ according to a reference machine are not significantly different.*

After 100,000 interactions the scores should be reasonably converged, making this a decisive episode length for testing the hypotheses. However, the speed of this convergence may also be affected by the manipulation of the action space. A supporting hypothesis for the strong interpretation can be stated as follows: *There is not a significant interaction between the episode length and the size of the action space among the group means of AIQ as well as SD of AIQ.*

3.2.2 SETTINGS

For BF 2, BF 10, and BF 20, 200,000 new environment programs were generated using the sampler by Legg and Veness (2013). The number of unique programs produced were 157,295, 160,609, and 160,619, for the BF 2, BF 10 and BF 20, respectively. Statistics according to program length are given in Table 1. Overall, the sample characteristics were fairly similar to those of the default case, with some differences in maximum length.

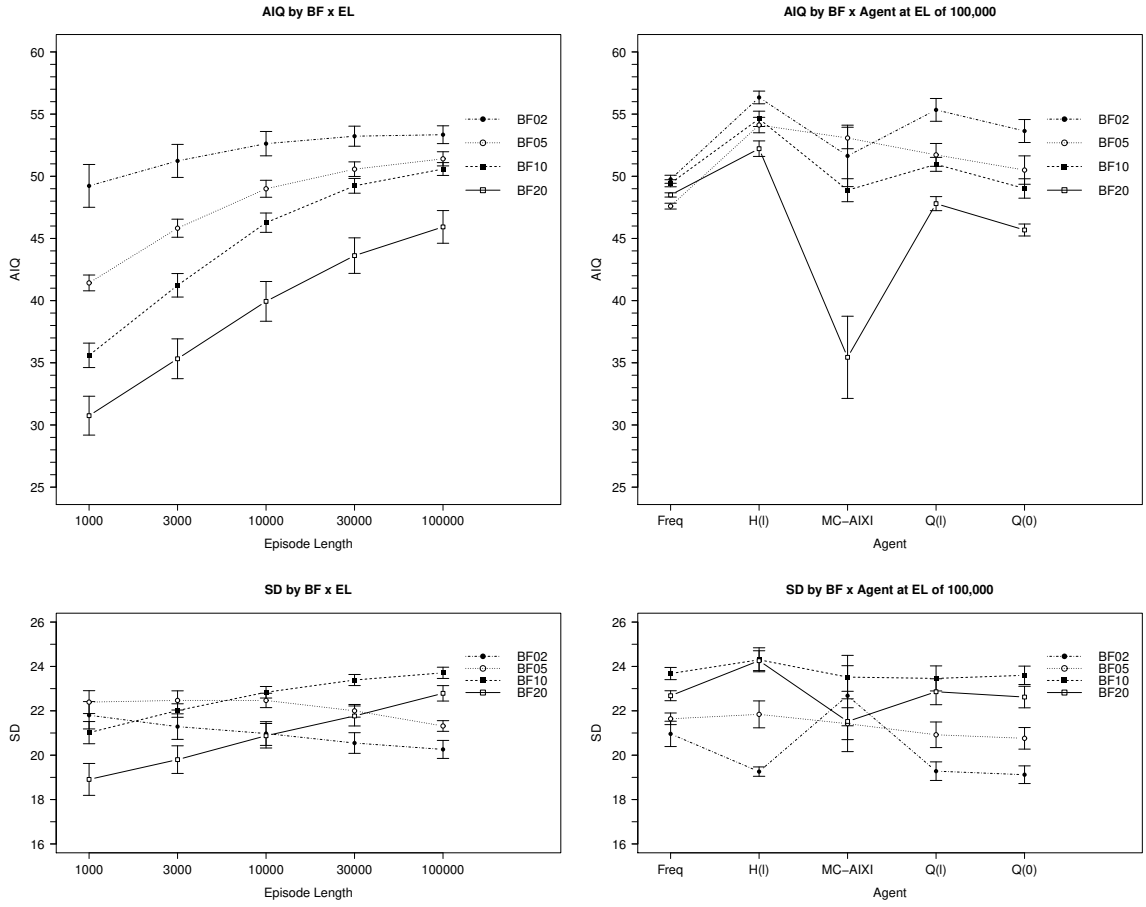

Figure 4: Interaction plots showing the impact of varying the action space on the means of estimated AIQ scores (top) and on the means of SD of estimated AIQ scores (bottom) of different episode lengths (to the left), and of tested agents at EL of 100,000 iterations (to the right). For *MC-AIXI* only configurations *#1 – #5* are used for the mean of BF 5.

Since no agent configuration for this experiment was provided by Legg and Veness (2013, 2011), the same settings were used as in Section 3.1. However, due to the computational requirements, only five configurations of *MC-AIXI* (*#1* to *#5*) were used. All other test settings remained the same.

3.2.3 RESULTS

The best achieved AIQ score estimates with a margin of error corresponding to a 0.95 confidence interval for each agent after the tested number of interactions are given in Appendix A, in Figure 9. Results of all tested agent configurations are shown in Appendix A, in Tables 9, 10, and 11.

3.2.4 DATA ANALYSIS

As can be seen in Figures 9 and 4, both the best achieved and mean results of agents differ from those achieve using BF 5 when the size of the action space is modified, changing the ordering of agents: Q_0 or even Q_λ perform in some cases as poorly as *freq* on BF 10 and BF 20. Some agents perform similarly for longer episodes (eg. HLQ_λ and Q_λ on BF 2). The AIQ scores of all agents except for *freq* seem to decline when the action space is increased, also the results are more spread out and the learning phase gets longer. *MC-AIXI* performs worse relative to other agents on BF 10 than in the default experiment, and in the case of BF 20 is in fact the worst. It seems to be affected particularly strongly by the action space adjustment, and this effect can be observed across all tested configurations.

Examining Figure 4, it seems that increasing the action space decreases the mean AIQ score at EL of 100,000 interactions. When broken down by agents, this effect is clearly a result of more than just the relatively poor performance of *MC-AIXI*. Also, there seems to be an interaction between the size of action space and the episode length when AIQ is concerned. This effect of increasing the action space can be seen to diminish with increasing episode length. Increasing the action space at an EL of 100,000 also seems to have an effect on the SD of AIQ score, albeit a more complex one. It seems to increase as the action space rises from 2-10 symbols, but decrease as the action space continues to rise from 10-20. By looking at the SD of AIQ of individual agents, it can again be seen that this effect on the overall score is not isolated to the performance of *MC-AIXI*. There also seems to be an interaction between the episode length and the size of action space when SD of AIQ is considered. While the SD of AIQ increases with increasing episode length for the 10 and 20 symbol action space, it decreases somewhat with increasing episode length for the 2 and 5 symbol action space.

To ascertain the significance of the effects and interactions observed in Figure 4, a statistical analysis was conducted using repeated measures ANOVA with sphericity corrections. For the analysis, the BF reference machine used is the manipulated within-subject factor with levels of BF 2, BF 5, BF 10, and BF 20 denoting the size of the action space. In cases where the interaction with episode length is considered, EL is an observed within-subject factor with levels of 1,000, 3,000, 10,000, 30,000 and 100,000 interactions. The subject is the tested agent configuration, giving 25 subjects to be tested at each factor level. Lastly, the dependent variable is the estimated AIQ score of the tested agent configuration at given episode length, or SD of the estimated AIQ respectively.

- According to this analysis, changing the action space size at an EL of 100,000 interactions did indeed have a statistically significant effect on the AIQ score of agents, $F(3, 72) = 20.98, GGe = 0.44, p = 1.9 \times 10^{-5}, ges = 0.30$.
- According to this analysis, changing the action space size at an EL of 100,000 interactions did indeed have a statistically significant effect on the SD of AIQ score of agents, $F(3, 72) = 34.57, GGe = 0.61, p = 2.4 \times 10^{-9}, ges = 0.42$.
- According to this analysis, there was, indeed, a statistically significant interaction between the episode length and the size of action space when the AIQ score of agents is concerned, $F(12, 288) = 30.86, GGe = 0.12, p = 1.8 \times 10^{-7}, ges = 0.059$.
- According to this analysis, there was, indeed, a statistically significant interaction between the episode length and the size of action space when the SD of AIQ score of agents is concerned, $F(12, 288) = 48.41, GGe = 0.29, p = 7.5 \times 10^{-20}, ges = 0.12$.

Longer runtimes were observed in accordance with the report of Legg and Veness (2013).

3.2.5 DISCUSSION

Based on the data analysis above, all hypotheses from Section 3.2.1 should be rejected:

- The ordering of agents according to their maximal and mean AIQ *is not the same* among the reference machine groups.
- Group means of AIQ according to a reference machine *are significantly different*.
- Group means of SD of AIQ according to a reference machine *are significantly different*.
- *There is a significant interaction* between the episode length and the size of the action space among the group means of AIQ as well as those of SD of AIQ.

Therefore, changing the action space indeed has an impact on the agents' results in the AIQ test.

Some caution is needed when interpreting the results, since the comparison is based on five configurations of each agent. Furthermore, the configurations were picked as a result of their strong performance on BF 5, which may not ensure the same level of performance when action space is modified, since it changes the complexity of the test and the configurations may be too finely-tuned to the complexity of BF 5 environments, effectively being curbed by that complexity. To fully ascertain the effects of varying the action space, a more complex but also more demanding set of experiments is needed. One featuring a full parameter sweep for all the tested agents. Nevertheless, there is a significant impact at least on the tested configurations of the tested agents.

Since the exact results and agent configurations used by Legg and Veness are unknown, the fact that this experiment achieved different results may not necessarily mean the test itself is unreliable. However, since the same configurations were used in this as well as in the default experiment it can be concluded that the test seems to be sensitive to changes in the size of the action space.

3.3 Varying the Observation Space

Another experiment mentioned by Legg and Veness (2013) increases the maximal number of observation symbols supplied to the agent from the default of 1. However, no details are given at all and results are only described as being "qualitatively the same" as the default case.

3.3.1 HYPOTHESES

Since a better formulated hypothesis is needed than just "results are qualitatively the same", the following specifications were devised, similarly to those described in Section 3.2.1.

- Weak interpretation – *The ordering of agents according to their maximal and mean AIQ is the same among the reference machine groups.*
- Strong interpretation – *Group means of AIQ according to a reference machine are not significantly different.*
- Supporting – *Group means of SD of AIQ according to a reference machine are not significantly different.*
- Supporting – *There is not a significant interaction between the episode length and the size of the observation space among the group means of AIQ as well as SD of AIQ.*

3.3.2 SETTINGS

BF 5 setups with 2, 3, and 4 observation symbols (BF 5,2, BF 5,3, and BF 5,4) were tried. Since BF 5,4 proved too time-demanding, it was dropped. For each setup 200,000 new environment programs were generated using the sampler by Legg and Veness (2013). The number of unique programs produced were 161,922, 162,189, and 162,238, for the BF 5,2, BF 5,3 and BF 5,4, respectively. Statistics according to program length are given in Table 1. Overall, the sample characteristics were fairly similar to those of the default case, with some differences in maximum length. The same agent configurations were tested as in Section 3.2. Also, other settings of the test remained the same.

3.3.3 RESULTS

The best achieved AIQ score estimates with a margin of error corresponding to a 0.95 confidence interval for each agent after the tested number of interactions are given in Appendix A, in Figure 9. Results of all tested agent configurations are shown in Appendix A in Tables 12 and 13.

3.3.4 DATA ANALYSIS

As can be seen from Figures 9 and 5, varying the observation space seems to have very little impact on both the best achieved and the mean AIQ scores of agents compared to the default experiment. The exception is *MC-AIXI* which achieves a slightly worse AIQ at best, only reaching the levels of Q_0, and Q_λ. Also, the configurations *#1* and *#3* perform especially poorly with BF 5,3, pulling the overall group mean down noticeably. Comparing their results in all experiments, it would seem that a *context tree depth* of 8 proves too restrictive for the agent in more complex settings (represented by BF 20, and BF 5,3), although this warrants a more thorough investigation.

Looking closer at Figure 5, at an EL of 100,000 interactions, increasing the observation space seems to have no effect on AIQ score. The slight decrease of AIQ for BF 5,3 seems to be caused by the poorer performance of some of the *MC-AIXI* configurations, as can be seen from the scores broken down by agents. There also seems to be no interaction between the EL and the observation space when AIQ score is concerned. The observation space increase may have some small effect on the SD of AIQ score at an EL of 100,000 that cannot be accounted for by the results of *MC-AIXI*, as well as some slight interaction of EL and observation space where SD of AIQ is concerned.

To ascertain the significance of the effects and interactions observed in Figure 4, a statistical analysis was conducted using a repeated measures ANOVA with sphericity corrections. In this case, the BF reference machine used is the within-subject factor with levels of BF 5,1, BF 5,2, and BF 5,3 denoting the size of the observation space. The rest of the parameters of the analysis are analogous to the action space variation experiment.

- According to this analysis, changing the observation space size at an EL of 100,000 interactions had no significant effect on the AIQ score of agents, $F(2, 48) = 1.66, GGe = 0.50, p = 0.21, ges = 0.043$.

- According to this analysis, changing the observation space size at an EL of 100,000 interactions did indeed have a statistically significant effect on the SD of AIQ score of agents, $F(2, 48) = 4.68, GGe = 0.59, p = 0.034, ges = 0.071$.

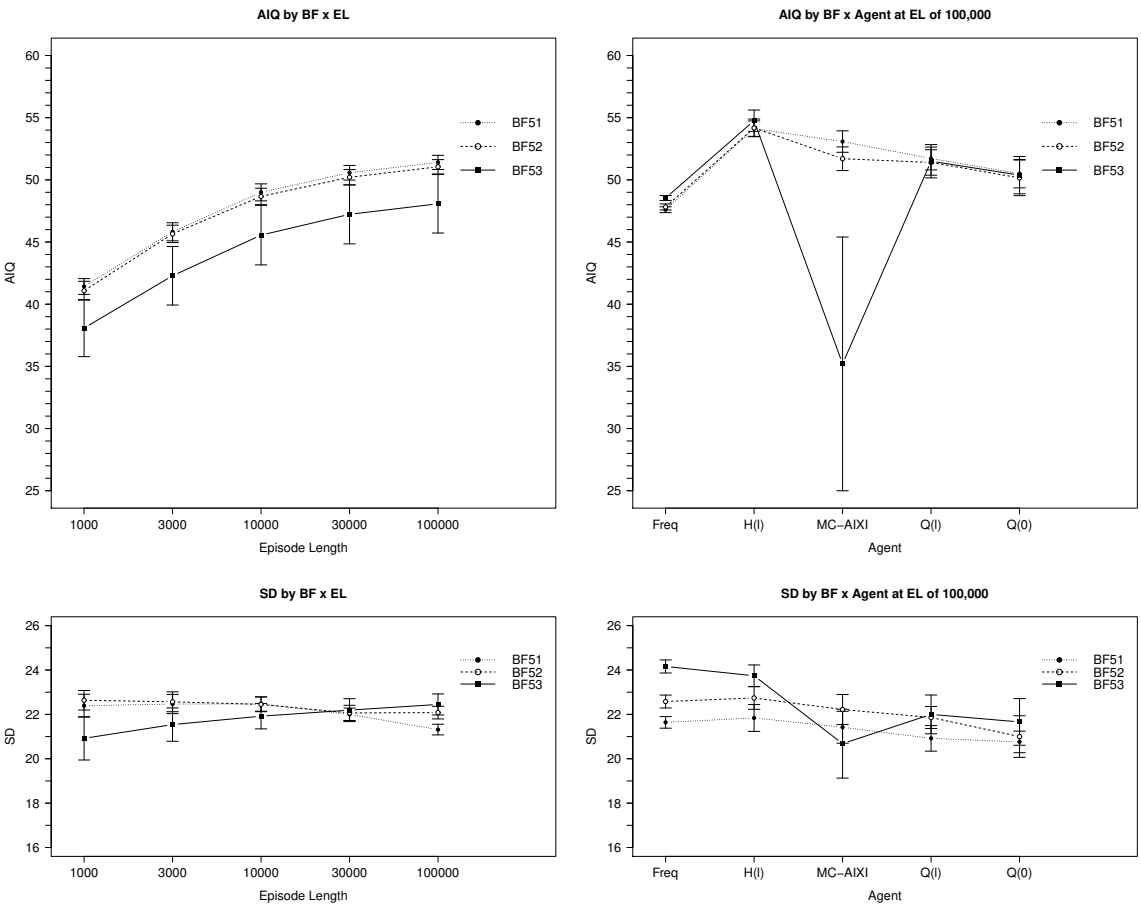

Figure 5: Interaction plots showing the impact of varying the observation space on the means of estimated AIQ scores (top) and on the means of SD of estimated AIQ scores (bottom) of different episode lengths (to the left), and of tested agents at EL of 100,000 iterations (to the right). For *MC-AIXI* only configurations *#1 – #5* are used for the mean of BF 5.

- According to this analysis, there was no statistically significant interaction between the episode length and the size of observation space when AIQ score of agents is concerned, $F(8, 192) = 0.21, GGe = 0.23, p = 0.79, ges = 5.8 \times 10^{-5}$.
- According to this analysis, there was a statistically significant interaction between the episode length and the size of observation space when SD of AIQ score of agents is concerned, $F(8, 192) = 24.45, GGe = 0.44, p = 1.9 \times 10^{-12}, ges = 0.027$.

Longer runtimes were observed in accordance with the report of Legg and Veness (2013).

3.3.5 DISCUSSION

Based on the data analysis above, some hypotheses from Section 3.3.1 were rejected:

- The ordering of agents according to their maximal and mean AIQ *is the same (with the exception of MC-AIXI)* among the reference machine groups.
- Group means of AIQ according to a reference machine *are not significantly different*.
- Group means of SD of AIQ according to a reference machine *are significantly different*, although the significance is weak and the effect size is small to medium.
- *There is not a significant interaction* between the episode length and the size of the observation space among the group means of AIQ, however, *there is a significant interaction* among the group means of SD of AIQ, although the effect size is small.

Therefore overall, it can be said that changing the observation space has no impact on the results of agents in the AIQ test.

When interpreting the results, some caution is needed since the comparison is based on five configurations of each agent. Furthermore, the configurations were usually picked so as to perform well on BF 5 which may not ensure the same level of performance when the observation space is modified, since it changes the complexity of the test and the configurations may be too finely-tuned to the complexity of BF 5 environments with single observation only, effectively being curbed by that complexity. Having said that, this only appears to be the case for two of the tested *MC-AIXI* configurations. To ascertain more thoroughly the effects of varying the observation space a more complex but also more demanding set of experiments using a full parameter sweep for all the tested agents is needed. Also, an investigation is needed into whether the complexity of all the environments increases as expected when the size of the observation space grows. Nevertheless, there is not a significant impact at least on the tested configurations of the tested agents.

Since the exact results and agent configurations used by Legg and Veness are unknown, the fact that this experiment achieved the same results may not necessarily mean the test itself is reliable. However, since the same configurations were used in this as well as in the default experiment it can be concluded that the test seems to be insensitive to changes in the size of the observation space.

4. Towards Addressing the Limits of the AIQ Test as a General AI Evaluation Method

Having tested the AIQ test practically, Section 4.1 contains observations about its suitability as a general AI evaluation method. Section 4.2 considers methods to address the observed limits of the test, some of which are demonstrated in Section 4.3. Finally, Section 4.4 shows ways of using a suitable general AI evaluation method.

4.1 Observations from Conducted Experiments and Related Work

In the previous section, an attempt was made to replicate the results given by Legg and Veness (2013, 2011). The main observation that can be drawn from the conducted experiments is that the AIQ test results are dependent on the parameters of the BF reference machine in a manner contrary to the original findings of Legg and Veness. While this seems not to be the case when varying the size of the observation space, there is a significant impact when the size of action space is changed which dominates the replication experiments.

It should be noted that based on the summarized related work, several concerns regarding the suitability of the AIQ test as a general AI evaluation method can be raised:

- As originally noted by Legg and Hutter (2007b) and further analyzed by Hibbard (2009) and Hernández-Orallo (2015), Universal Intelligence, and therefore the AIQ test results, depend on the choice of reference machine. The observed dependence of the test results on the action space size of the BF reference machine seems to be in support of this concern.

- The Universal Intelligence definition considers agent's performance in all computable environments. These necessarily include, as Hernández-Orallo and Dowe (2010) correctly point out, many non-discriminative environments including so called "heaven" or "hell" environments. While this issue might not be so critical when looking for a definition, it is of crucial importance when discussing a practical test for which there are only limited resources available (which would be wasted by these non-discriminative environments) and only a sample of environments is tested (which can be skewed by the non-discriminative environments). The AIQ test is not guarded against such cases unlike the AIT.

- The Universal Intelligence measure, as well as the derived AIQ score, only implicitly contains some aspects of intelligence, while others are not reflected at all. While the score integrates the measure of an agent's success in environments and the measure of its generality, coined as Intellectual Breadth by Goertzel (2010), neither of these are explicit which makes it difficult to form a detailed agent comparison. Further, the aspect of effectiveness is not considered even though it is advocated e.g. by Goertzel (2010), neither is the aspect of time though proposed by Hernández-Orallo and Dowe (2010).

Furthermore, based on the conducted experiments, the following observations regarding the test complexity can be made:

- Given the test format, computational requirements of an agent become apparent. This is due to the high number of agent–environment interactions necessary for a sufficient AIQ score convergence, as well as due to the high number of environment programs necessary for a reasonable representativeness of AIQ estimate.

- For AIQ score to converge, many agent–environment interactions are needed. Their number differs both among the agents as well as among the BF reference machine settings.

- Response times of an agent can differ during the test episode.

- Different configurations of the same agent can score rather different values of AIQ. The configuration space of an agent can be large.

- Since the differences in scores among tested agents are rather limited, the default test setting may still be too simple.

Any evaluation process based on the AIQ test should be devised with these observations in mind. However, these observations are of concern for any other similarly general and complex test involving many interactions with many environments. In fact, this is also the case for the Anytime Intelligence Test, although its adaptive nature reduces some of the concerns.

To allow for more observations and better interpretation of the test results, a deeper analysis of the AIQ test's inner workings is needed. This could be achieved by looking closer at the environment programs, focusing on the way in which rewards and observations are computed. A preliminary investigation indicates that in some environments rewards or observations are solely dependent on chance, while in others they depend on an agent's actions. This suggests that it is feasible to classify environment programs according to some criteria and further investigate the impact of such criteria on an agent's AIQ. While this investigation may focus on the role of chance and an agent's actions in computing rewards and observations, other aspects could be analyzed, including the role of the complexity of an environment program, pointless code, or premature termination of environment program execution. The method, further referred to as semantic analysis of environment programs, consists of identifying the semantics of an environment program class and describing its possible syntax in BF language using regular expressions. While such approach is necessarily incomplete as all possible syntactical combinations for a given semantics cannot be listed, practically, it is possible to capture a large proportion of environments in question. Preliminary results from this analysis will be used in the next section to illustrate the prevalence of the addressed issues.

4.2 Proposals of AIQ Test Improvement

As several concerning observations were made in the previous section, this section will discuss them in more depth and propose possible ways to alleviate the issues.

4.2.1 REDUCING DEPENDENCE ON REFERENCE MACHINE

The bias of the Universal Intelligence measure towards a certain class of environments due to the choice of reference machine can be arbitrarily reduced by specifying a minimal length for environment programs, as suggested by Hibbard (2009). The bias can also be reduced by changing the way an agent's overall score is computed to be based on the concept of difficulty, as suggested by Hernández-Orallo (2015).

Let us first examine the proposal of Hibbard (2009). As for the AIQ test implementation by Legg and Veness (2011), the minimal length of a program is 3 instructions, since read, write, and end program instructions have to be present in every program. Furthermore, the adaptive stratified score estimation procedure reduces the proportion of programs with fewer than 10 instructions in the sample.

Figure 6 compares empirical cumulative distributions of program lengths to illustrate how successfully the problem is dealt with in the current version of the AIQ test:

- The version of BF language used by the test has very few syntactical limitations, resulting in a cumulative distribution shown by the dotted line. Here, short programs dominate the sample with 44 % featuring a length of up to 3 instructions and 75 % with a length of up to

10. Also, programs of length 1 and 2 are syntactically valid, and many of the programs do not do anything useful.

- The BF sampler used by Legg and Veness (2011) imposes further restrictions on environment programs (namely interactivity, computation limit, and some code optimisations) resulting in a cumulative distribution shown by the dashed line. Here, short programs are less dominant with about 3 % featuring a length of up to 3 instructions, and 38 % up to 10.

- However, the AIQ test uses adaptive stratified sampling to reduce the variance of the AIQ estimate, therefore the actual distribution of tested programs differs from the distribution in the sample. Also, there are differences in the distributions among agents based on their results (further influenced by episode length). A cumulative distribution resulting from a total of 20 test runs of EL of 100,000 interactions with configurations *#1* to *#5* of agents *freq*, Q_0, Q_λ, and HLQ_λ is shown by the solid line. Here, short programs are even less dominant with only about 1 % featuring a length of up to 3 instructions, and 18 % up to 10.

Implementing Hibbard's restriction should be relatively straightforward, although some changes to the environment program stratification procedure are needed. The strata decided on according to program length would have to use updated limits reflecting the new minimal length of a program. The strata decided on according to an exhibited simple pattern in returned rewards might actually have their criteria changed since the probability of a program exhibiting the given pattern may become too low for high minimal lengths.

However, the main concern regarding Hibbard's restriction is that it raises the question of defining a suitable minimal length setting. The number of distinct instructions of a given reference machine is a potentially interesting limit, since shorter programs necessarily miss some of the instructions and are, therefore, somewhat syntactically degraded. Given their short length, the syntactic degradation cannot be compensated for (some of the instructions can only be shorthands for complex expressions), and becomes semantic. This limit can still be rather weak for very simple reference machines (such as the BF used in the AIQ test), however, it can also be too strong for very complex reference machines (such as some high level programming languages). It could even undermine the very idea of the Universal Intelligence definition, that the entity has to be explicitly tested in simple environments, not only in complex ones. As such, it has to be approached carefully.

With all that considered, implementing Hibbard's length restriction as a free parameter of the AIQ test could spark interesting empirical research into the dependence of the score on the reference machine, and using the number of distinct instructions can be considered a reasonable lower limit of the minimal length restriction. For an AIQ test running on the BF reference machine, this would mean setting the limit to 10 instructions, effectively replacing 38 % of the current sample with longer programs.

The proposal of Hernández-Orallo (2015), on the other hand, requires more extensive changes to the AIQ test. There seem to be two approaches to incorporating the difficulty-based evaluation into the current test:

1. The stratification procedure can be modified to stratify based on the difficulty estimation of the generated environment programs. At the same time, the resulting distribution of the program difficulty strata have to be kept either uniform or slowly decaying. Then the overall score produced by the test would be an estimate of the score proposed by Hernández-Orallo (2015).

2. If the environment programs actually used in the test are saved and paired with the agent's results, these can then be used to estimate the actual difficulty of environment programs through the construction of agent response curves. With that an estimate of the score suggested by Hernández-Orallo (2015) can be computed independently.

The choice between the options 1 and 2 depends on how closely the difficulty estimation process should be tied to the actual agent being tested. While the practical implementation is not as straightforward as in the case of Hibbard's length restriction, this proposal of Hernández-Orallo (2015) shows inspiring direction for future work.

4.2.2 DEALING WITH NON-DISCRIMINATIVE ENVIRONMENTS

The AIQ test is not well-guarded against non-discriminative environments by testing only with reward-sensitive environments as Hernández-Orallo and Dowe (2010) suggest. Environments that stop interacting or that interact for too long are excluded due to the mandatory read and write instructions as well as the imposed computation limit. Purely "heaven" or "hell" environments (those that give constant rewards all the time) are also excluded by Legg and Veness (2013, 2011). However, environments featuring potential "heaven"/"hell" situations (having a "heaven"/"hell" subenvironment accessible on some conditions), as well as environments with random rewards or observations remain.

A preliminary semantic analysis of environment programs indicates that in the case of the BF 5 sample, about 17 % of environment programs surely result in an agent's actions having no effect on its rewards. These are cases in which the resulting reward is random. Also, the mitigation of the mentioned problems with environment programs is not always complete.

There are basically two possible solutions for the problem of non-discriminative environments in the AIQ test:

1. Switch to a reference machine with a proven ability to generate only reward-sensitive environments while still being universal. Such a machine was defined and its properties were proven by Hernández-Orallo (2010). However, it has not yet been fully implemented, since Insa-Cabrera et al. (2011) used only a simplified, non-universal version. Further, this reference machine also solves the issue of balanced environments which is already solved in a different way by the AIQ test, so an additional modification of one or the other would be needed.

2. Decrease the proportion of problematic environments in the BF reference machine sample. Some of the non-discriminative environment programs have an easy-to-describe pattern either in their source code (as is shown by the preliminary semantic analysis) or in the interaction sequence they create (e.g. return constant reward from a certain point in interaction sequence). While such an approach would not mitigate the issue completely, it may reduce it reasonably.

However, it could be argued that the required property of reward-sensitivity as formulated by Hernández-Orallo and Dowe (2010) is too strong. Namely, Hernández-Orallo and Dowe require that "... at any point/time there are *always* at least two different sequences of actions that can lead the agent to get different accumulated rewards for n interactions" (italics added for emphasis). While the environments that only produce the problematic events (random/constant rewards all the time) are obviously useless, the environments that produce the problematic events only on some

conditions (e.g. only after a certain action following a certain observation) can have some evaluation purpose. That is, the agent should be enabled to "fall in a hole and die" especially if the hole is properly advertised by observations. Actually, the requirement explained by Hernández-Orallo and Dowe as "...environments can have an agent stuck for a time (in a 'hole') if the good actions are not taken, but there is a way to get out of there or at least to find different rewards inside the hole" might be counterproductive since it effectively discourages self-preservatory behavior. Therefore, detectable and avoidable (even and especially if not by all agents) traps should be allowed. So perhaps, instead of the original strong formulation, some week reward-sensitive environments could be of interest.

With that considered, improving the BF sampler of the AIQ test to exclude more of the reward-insensitive environments (in the weak formulation of the concept) and running a comparative group analysis against the original sampler on the test results may give interesting empirical results regarding the real extent of the problem with non-discriminative environments.

4.2.3 INVESTIGATING POSSIBLE SIMPLICITY OF THE TEST SETTING

Both the best achieved and average scores among the tested agents show rather limited differences and are placed around the middle between the minimal and maximal AIQ values. This is supported by the original results of Legg and Veness (2013) as well as the replicated results in this paper (with added details). The test may well be valid and the differences shown may be real, however the test setting could also be too simple or the way the score is aggregated may favour simple environments too greatly, thereby failing to bring out the real differences among agents.

Finding the truth among these possibilities is not easy, but the following methods can help to answer the question:

- Observe the changes in differences between agent scores when the complexity of the test is increased.
- Compare the differences between agent scores achieved in the AIQ test with the differences in other general tests.
- Analyze the differences between agent scores in environments of varying complexity featured in the AIQ test.

What could the results of the outlined investigations be and what would they mean for the issue at hand? Basically, either rather similarly limited differences would be observed, or the differences would (profoundly) change. The evaluation of this question needs to be sufficiently fuzzy, especially in case of comparison to other (possibly very different) tests. If the differences remain sufficiently alike in other scenarios, it would suggest that the differences are real. However, if the differences change extensively, it would suggest that the differences are only due to the test setting simplicity.

While this ideally calls for a new set of experiments focused on increasing the test complexity and eliminating the possibility of agent configurations being too finely-tuned towards a certain test setting by doing a thorough parameter sweep, the results of conducted experiments and the overviewed literature, could (with some caveats) serve as an illustration of the outlined approaches.

Experiments conducted in Sections 3.2 and 3.3 should increase the complexity of the test setting by increasing the size of action space (and the derived number of symbols used by the BF reference machine), or by increasing the size of observation space respectively. The interaction

Table 4: Results of agents per stratum averaged from results of configurations *# 1–5* weighted by the number of tested programs for a given stratum on episode length of 100,000 interactions (BF 5 Reference Machine).

Agent	**Average Rewards per Stratum**									
#	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
%	*17.6*	*10.4*	*2.5*	*0.6*	*1.9*	*1.8*	*1.8*	*1.8*	*1.8*	*0.7*
freq	92.9	78.0	75.0	73.0	93.0	93.0	79.5	77.6	94.0	78.6
Q_0	97.3	86.1	76.4	64.0	97.1	97.2	87.9	89.9	96.1	88.2
Q_λ	97.3	91.5	80.1	70.5	96.9	96.7	92.5	92.0	96.2	92.0
HLQ_λ	97.7	97.0	93.7	92.0	97.5	96.6	97.0	96.9	96.0	96.0
MC-AIXI	98.9	97.4	81.8	80.1	99.0	99.0	98.6	92.1	96.6	87.0
#	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
%	*5.9*	*4.6*	*6.8*	*6.2*	*6.9*	*5.5*	*6.8*	*6.9*	*4.7*	*4.8*
freq	0.0	20.7	24.9	22.7	21.3	24.2	22.7	22.4	25.1	23.7
Q_0	0.0	21.4	27.5	24.6	21.8	24.3	23.7	23.6	25.2	23.4
Q_λ	0.0	22.3	27.9	25.8	22.5	25.2	24.7	24.5	26.7	24.9
HLQ_λ	1.6	24.2	30.0	27.6	24.0	27.7	27.2	27.3	29.0	28.0
MC-AIXI	0.0	17.1	33.7	28.4	19.9	15.9	27.0	26.6	34.3	24.6

plots in Figures 4 and 5 suggest that there is in fact a change in the differences among the scores of agents when the action space is varied, while there is no change in the score differences when the observation space is varied. This discrepancy may, however, only be apparent, since the preliminary semantic analysis of the environment programs indicates that in about 31 % of the environment programs no observation is produced in computation and therefore increasing the size of observation space in those environments cannot have any effect on the complexity. Furthermore, only in 9 % of the environment programs, more than one observation is always produced during the computation. Therefore, the results from the experiment varying the action space dominate.

Insa-Cabrera et al. (2011) conducted an experiment with a simplified version of the Anytime Intelligence test comparing humans to a Q-learning agent. It might be informative to conduct a test using their settings and the agent configurations tested in this paper, however this would be rather demanding since some interfacing of python and C-based agents to a Java testing platform would be needed. The test would also have to be run using many agent configurations and environments, which is beyond the scope of this paper. However, comparing the published results of Insa-Cabrera et al. with the results of this paper could still be useful. Regarding the question at hand, the important result of Insa-Cabrera et al. (2011) is that they did not find a significant difference between the tested Q-learning and human intelligence, suggesting that the test setting they used is too simple to bring out the expected difference. Since the AIQ test has not been used on humans (although it is possible to input actions manually and the interface approach of Insa-Cabrera et al. shows a promising method to make it user-friendly), it cannot be compared directly with results of Insa-Cabrera et al. However, since the AIQ test shows some difference among the tested agents, the test setting of Legg and Veness (2013) is probably not as simple as the setting of Insa-Cabrera et al. (2011).

Legg and Veness (2011) stratify environment programs into 20 strata based both on the presence of simple patterns in returned rewards (strata 1–10), and the program length if the simple patterns

are absent in returned rewards (strata 11–20). The test can print the results of an agent in individual strata. An analysis of the differences among the agents can then attribute these differences to a certain stratum, or at least to a group of strata with simple environment programs (strata 1–10) or more complex ones (11-20). Looking at the agents' results per stratum in Table 4, it seems that the strata can indeed be partitioned into the previously mentioned groups based on complexity. However, stratum 11 most likely contains programs returning either close-to-random or close-to-zero rewards constantly. Focusing on the titular question of this section, there are indeed changes in differences of agents' results per stratum. However, a quick inspection of the differences does not show the expected pattern between the two groups. With the pattern of differences being possibly more fine-grained, a deeper analysis (which would greatly benefit from implementing the suggestions of Section 4.2.2) is required in order to get more conclusive results.

The preliminary results, based on the current level of detail of the analysis, are therefore inconclusive. As such, a more thorough investigation into the potential effects of an overly simplistic test setting is needed, following not only the outlined approaches but also additional perspectives.

It is also possible that the perceived limited differences among agents are only apparent. While the AIQ may invite a linear interpretation as a result of its formula, the test is based on the Universal Intelligence, which uses exponential weighting by environment complexity in the form $2^{-K(\mu)}$. This is approximately kept in the way the environment programs are sampled by the AIQ test. Thus, a logarithmic interpretation of the AIQ score may not be improper, at least in cases when the difference in scores is caused by a success in an additional more complex environment. However, it is currently unclear to what extent a logarithmic interpretation would pose a problem in cases where the difference is solely due to an improved performance in the same environments, or due to a mixed case. Supporting the AIQ score by additional characteristics (as discussed in Section 4.2.4) may help, since in the current AIQ test it is difficult to tell the precise source of the difference.

4.2.4 MEASURING SUPPORTING QUALITIES OF INTELLIGENCE

The AIQ score combines a measure of an agent's success in an environment along with a measure of its generality, however it does not include a measure of an agent's time-frame as advocated by Hernández-Orallo and Dowe (2010), nor a measure of its computational effectiveness as suggested e.g. by Goertzel (2010). While these additional aspects can be integrated into an overall score, as the authors suggest, it would make interpreting the test results even more complicated.

In the current version of the test, it is actually difficult to tell whether a change in AIQ is due to the change in the agent's ability to achieve goals, or to the change in its generality. Integrating the dimension of time or computational effectiveness would be at the cost of adding more dimensions of uncertainty when interpreting the results. While it can be argued that the aspects of success and generality are sufficient for the definition of intelligence as an ideal concept, it is also true that the aspects of effectiveness and time-frame are of at least practical concern.

Instead of integrating further aspects into the overall score (or perhaps besides integrating them), the following measures should accompany the AIQ score to characterize the agent in detail:

- A measure of the agent's generality which could be based on a modified and completed version of Intellectual Breadth by Goertzel (2010).
- A measure of the agent's time-frame which could be based on the time-sensitive formalization of the Anytime Intelligence Test by Hernández-Orallo and Dowe (2010) as well as a measure of the dynamics of agent's response times.

- A measure of the agent's computational effectiveness as proposed e.g. by Goertzel (2010).
- A measure of the AIQ score convergence speed as discussed in later section of this paper.

Having the AIQ score augmented by these accompanying measures would facilitate both the comparison of different agents as well as an improved version of the same architecture.

4.2.5 DECREASING HIGH COMPUTATIONAL REQUIREMENTS

The format of the AIQ test highlights the computational requirements of tested agents. It can be controlled by setting several parameters of the test.

In the case of episode length and environment programs sample size, an increase in computational requirements leads to more precise results. A testing procedure based on a concept of sufficiently precise results can be adopted, if suitable stopping criteria are found and set. Such a procedure would include several testing rounds progressively increasing the episode length/sample size respectively, and evaluating for each agent configuration if the stopping criterion is met. Only the configurations not reaching a stopping criterion advance to the next round. Adopting this testing procedure, the demanding test runs with long episodes or large sample sizes are used only if they can meaningfully contribute to an increase in the precision of the results. An example of possible stopping criteria includes absolute or relative difference in AIQ scores, or a statistical test of significant difference of two means. The next section will elaborate on another criterion based on convergence speed.

As for the sizes of action and observation spaces of the BF machine, the increase in computational requirements should lead to the increase of test complexity. It is unclear, whether or not there is a preferable setting for these parameters. However, in the case of a suspected difference in the intelligence of two agents, a testing procedure could be attempted which would gradually increase the action and observation spaces, until the either the difference between agents shows, or further continuation becomes infeasible due to computational requirements. Again, such an approach could decrease the computational resources required when compared to conducting tests on a full predetermined set of BF reference machine settings.

On the more practical side of things, it would be beneficial to be able to resume incomplete test runs as well as combine results of a given environment program sample size to increase the precision of AIQ estimates. Also, the ability to save results of a tested agent after some intermediate number of interactions would speed up the testing process, since currently results for different episode lengths need to be computed separately. In the conducted experiments, each agent configuration was therefore tested in a total of 144,000 interactions instead of 100,000 as enabled by the proposed improvement. A similar change could be considered for results at intermediate sample sizes if integrated with the proposed multi-round method for sufficiently precise AIQ estimation.

4.2.6 ADAPTING TO DIFFERENCES IN NUMBER OF INTERACTIONS TO CONVERGE

Since the AIQ test is based on the Universal Distribution, the AIQ score of a tested agent should converge. Practically, the concept of a sufficiently converged AIQ score is of more interest. Also, the convergence is influenced by both the BF reference machine settings as well as by the configuration of an agent. Therefore, using a predetermined episode length as a test setting is ineffective, since in some cases the AIQ may be sufficiently converged already, while in other cases the score may not be sufficiently converged yet.

To address the issue, a measure describing the convergence process is needed. One such measure is the current rate of convergence $v_c^{\hat{\Upsilon}} = \frac{\Delta\hat{\Upsilon}}{\Delta T}$, where $\Delta\hat{\Upsilon}$ denotes the change in the AIQ score that occurred with the change in the number of interactions denoted as ΔT. This measure shows an average change in AIQ score per interaction computed from some interval of the agent – environment interaction sequence. For convenience (due to the numerical disproportions) it may be shown in a form per thousand interactions.

The current rate of convergence can be used as a stopping criterion for the proposed multi-round method for sufficiently converged AIQ score. More precisely, having a modified AIQ test so that it saves the current AIQ estimate after every n interactions, and a (reasonably set) minimal current rate of convergence, the method would compute the current rate of convergence from the two AIQ estimates and only continue the test if said rate is above the set minimum. Using this method, the test run would be terminated once the sufficiently converged AIQ is achieved resulting in a reasonably short episode length.

4.2.7 TESTING AGENTS WITH LARGE CONFIGURATION SPACE

Some agents have a large configuration space. As such, their performance in an intelligence test can vary greatly depending on their configuration. Given evenly-picked configurations or a random sample from the whole configuration space, a simple descriptive statistics can capture the differences.

However, given prior knowledge about the agent, certain regions of the configuration space can be searched more thoroughly while others can be mostly ignored, skewing the simple statistics in favor of the better mapped regions. This calls for a system of weights representing the coverage of the configuration space tested. The issue can actually be further divided into two cases:

- difference in the number of configurations tested for each value of a parameter,
- imbalance in coverage of the parameter value space by the choice of tested values.

To compensate for a difference in the number of configurations tested for each value of a parameter, value group means of other parameters should be constructed from value group means of the imbalanced parameter and not directly from the data.

To compensate for an imbalance in coverage of the parameter value space, weights should be assigned to parameter values depending on the parameter type:

- *Boolean* – can be simply weighted in ratio $1 : 1$.
- *Category* – can also be weighted equally according to the number of categories.
- *Bounded real number* – can be weighted in proportion to the interval the number represents. The endpoints of the interval can be either a mean of two chosen values of the parameter closest to each other, or the endpoint of the parameter interval in case there is no other chosen value in between them.
- *Natural number*:
  - In cases where a reasonable upper limit can be found, e.g. when it is computationally infeasible to test with higher than a certain value, natural number parameter values can be weighted as bounded real numbers.

– If no reasonable upper limit can be found, parameter values can be weighted in proportion to the differences of the consecutive tested values. The weight of the lowest value can be its difference from 0.

Having a set of weights w_i assigned as suggested above, a parameter coverage mean AIQ $\bar{\Upsilon}_C$ of an agent π according to its parameter p with n tested values can be computed as a weighted mean of the parameter value group means $\bar{\Upsilon}_G$ as shown by equation 3. This equation is in fact a special case of equation 1 by Legg and Hutter (2007b) with a particular reference machine.

$$\bar{\Upsilon}_C(\pi_p) = \sum_{i=1}^{n} w_i \bar{\Upsilon}_{G_i}(\pi_p) \tag{3}$$

4.3 Demonstrations of AIQ Test Improvements

As several ways to improve the AIQ test were proposed in the previous section, this section will briefly demonstrate some of them. See Appendix C for details on the extended AIQ test.

4.3.1 REDUCING DEPENDENCE ON REFERENCE MACHINE

Based on the discussion in Section 4.2.1, the sampling procedure of the AIQ test was extended so that it was possible to specify the minimal program length. Two methods were tried: 1) extending sampled programs of insufficient length until the condition was met, and 2) dropping programs of insufficient length until a program meeting the condition was sampled. The stratification procedure now takes into account the minimal program length when stratifying programs by length. The stratification according to an exhibited simple pattern in returned rewards was not modified.

A comparison of the new methods with the original sampling procedure is shown in Table 5. While the original BF 5 sample contains 40.9 % of simple programs (strata 1 – 10), setting minimal program length to 10 instructions reduces their proportion to 34 % using method 1) and even to 25.7 % using method 2). Figure 6 compares empirical cumulative distributions of program lengths. Among the samples with a minimal program length of 10, there is a higher probability of a program being syntactically valid and interactive than in the original sample. Also, since the longer programs are more diverse, giving higher variance in agents results, the benefit of using an adaptive sampler is reduced. Method 2) favors longer programs compared to method 1). The method that drops shorter programs should, therefore, be preferred over the method that extends shorter programs.

A statement by Hibbard (2009) regarding the reduction of the dependence on reference machine can now be practically validated by repeating the experiment that manipulates the size of the action space while also changing minimal program length. The impact of action space size manipulation should diminish with the increasing minimal program length. As the experiment is demanding, it is out of the scope of this paper. However, it will be attempted as a part of future work.

4.3.2 DECREASING HIGH COMPUTATIONAL REQUIREMENTS

In order to make the AIQ test more effective, a method proposed in Section 4.2.5 was implemented that saves the current AIQ score after every 1,000 interactions. Thus, the experiments' settings reported in Section 3 can be computed with only 2/3 of the required resources. Using this option a more detailed set of results documenting the score convergence process can be obtained, as shown

Table 5: A comparison of distributions of environment programs in strata among the original program sample, and samples with minimal program length of 10 instructions generated by the method 1) (extended), and 2) (dropped). (BF 5 Reference Machine)

Sample	**Strata Proportions**																			
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
original	17.6	10.4	2.5	0.6	1.9	1.8	1.8	1.8	1.8	0.7	5.9	4.6	6.8	6.2	6.9	5.5	6.8	6.9	4.7	4.8
extended	13.6	7.8	3.0	1.0	1.8	1.9	1.5	1.5	1.3	0.7	28.3	4.4	5.4	4.3	4.5	3.5	4.3	4.6	3.2	3.6
dropped	10.6	5.4	1.8	0.6	1.3	1.3	0.9	1.0	2.0	0.9	25.7	5.0	6.4	5.3	5.8	4.4	5.8	6.2	4.5	5.2

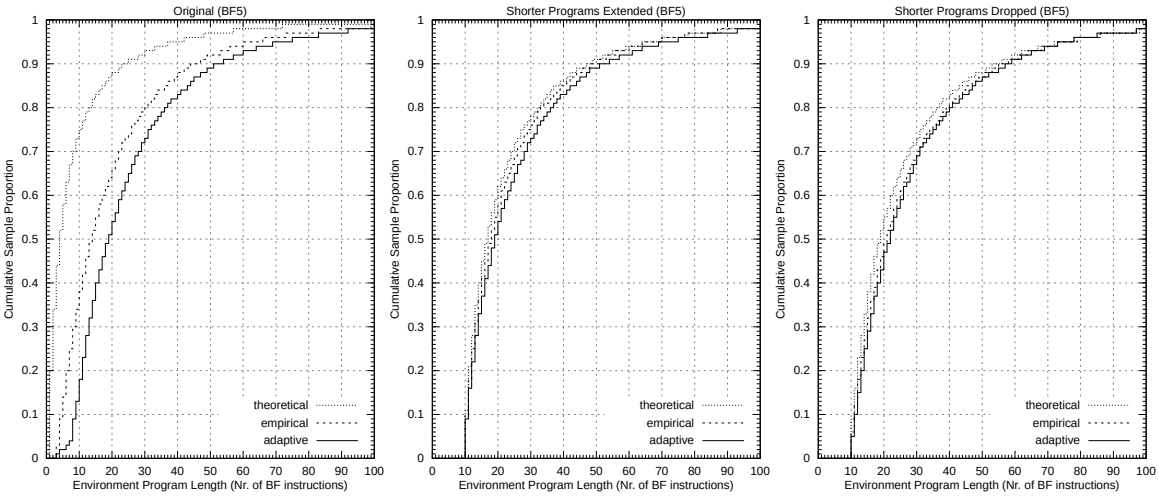

Figure 6: A comparison of cumulative proportions of BF 5 environment program lengths (up to 100 instructions): all syntactically valid BF programs (theoretical), programs following the restrictions of Legg and Veness (2013, 2011) (empirical), programs actually chosen for 20 test runs (adaptive). The programs sampled by the original BF sampler are to the left, the programs with minimal length of 10 sampled by extending shorter programs are in the middle, and the programs with minimal length of 10 sampled by dropping shorter programs are to the right.

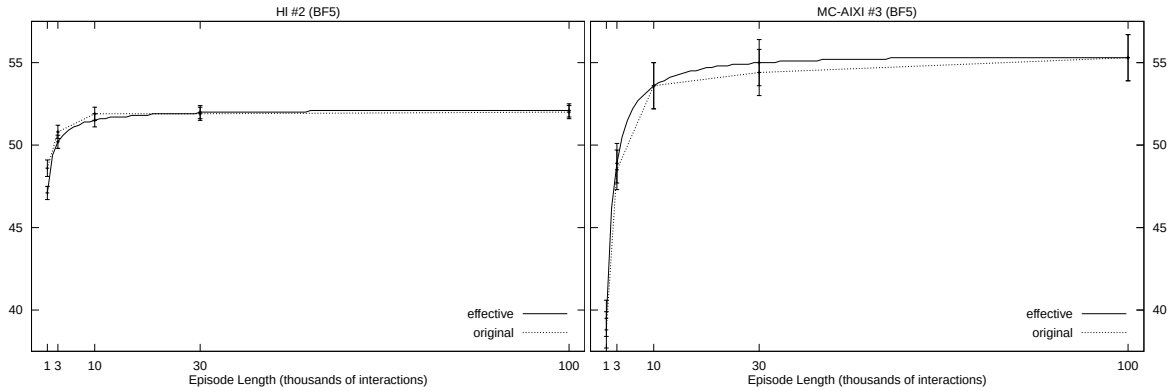

Figure 7: A comparison of an agent's AIQ score convergence process on BF 5 reference machine as obtained by the original test settings (original) and by the improved test settings (effective).

by Figure 7. Moreover, implementing this method facilitates further improvements proposed in Section 4.2.5.

To validate the implemented method, an experiment was conducted according to the default settings as described in Section 3.1. The results were computed directly for 100,000 interactions, with intermediate results saved after every 1,000 interactions. The new results were compared to the results reported in Table 7. Two-sample t statistics were computed to determine the significance of the differences of the agent results both at the final EL of 100,000 interactions, as well as at the intermediate lengths of 1, 3, 10 and 30 thousand interactions.

As for the EL of 100,000 interactions, the test did not reject that the difference is zero. At shorter episodes (mostly at 1,000 and 3,000 interactions), however, there were a total of 16 cases when the test rejected that the difference is zero. Since the largest difference was -1.5 ± 0.6, the implemented method can be considered valid. The method gives slightly lower AIQ estimates for shorter episodes, most likely due to the adaptive sampler now choosing programs to minimize the variance at 100,000 interactions when the agent had better learned the regularities of the environment.

4.3.3 TESTING AGENTS WITH LARGE CONFIGURATION SPACE

The following example, based on a parameter sweep of *MC-AIXI* in the default setup, should illustrate the method that deals with a large configuration space of agents as proposed in Section 4.2.7.

There was a difference in the number of configurations tested with and without *exploration decay*, therefore value group means of other parameters must compensate for this. In the case of *exploration*, 4 values were only used in combination with *exploration decay* while 4 other values were only used with no *exploration decay*, canceling the need to compensate on the level of group means. Also, the following parameters of *MC-AIXI* had their value space searched unevenly: *context tree depth* (natural), *exploration* (bounded real), and *exploration decay* which can be viewed either as a bounded real, or as a Boolean compound of using (bounded real) and not using exponentially decayed exploration. Considering that other methods of decaying the exploration strategy might be implemented, thus creating an additional category attribute, this would seem to

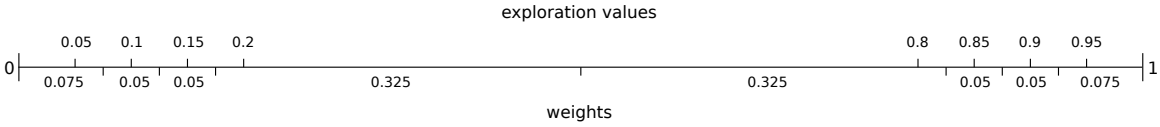

Figure 8: Intervals represented by parameter values and their assigned weights for the *exploration* parameter of *MC-AIXI*.

Table 6: The computation of parameter coverage mean AIQ scores $\bar{\Upsilon}_C$ (*) of *MC-AIXI* from value group means $\bar{\Upsilon}_G$ (AIQ) for an episode length of 100,000 interactions.

	MC Simulations			**Context Tree Depth**			**Search Horizon**			**Exploration**			**Exploration Decay**		
#	*value*	*weight*	*AIQ*	*value*	*weight*	*AIQ*	*value*	*weight*	*AIQ*	*value*	*weight*	*AIQ*	*value*	*weight*	*AIQ*
1	50	0.5	43.55	8	0.25	45.21	1	0.2	36.48	0.05	0.075	45.12	0.3	0.225	47.68
2	100	0.5	45.52	16	0.25	44.22	2	0.2	46.68	0.1	0.05	41.91	0.6	0.15	48.2
3				32	0.5	44.17	3	0.2	47.95	0.15	0.05	38.59	0.9	0.0875	48.91
4							4	0.2	45.91	0.2	0.325	35.78	0.95	0.0225	48.79
5							5	0.2	45.63	0.8	0.325	48.91	0.99	0.01125	49.33
6										0.85	0.05	48.81	0.995	0.00375	49.38
7										0.9	0.05	48.6	1	0.5	40.35
8										0.95	0.075	48.52			
*			44.53			44.44			44.53			43.44			44.25

be a preferable approach. The assigning of an interval represented by the parameter value and its weight is shown in Figure 8 for the example of *exploration*. The resulting weights of all parameters, the respective value group means for 100,000 interactions, and the parameter coverage means are summarized in Table 6. In the given example, compensating for the unequal number of tested configurations for each parameter value has a substantial impact on the resulting parameter coverage means. The differences caused by parameter value space coverage imbalance are less pronounced.

4.4 Using a Suitable General AI Evaluation Method

Having a suitable general AI evaluation method, there are still two important questions: "How to use it?" and "What to use it for?"

The 'how' question is tightly related to the observation of differences among configurations of an agent (or among individuals of a species). This calls for a statistics-based approach in presenting and evaluating the results (box plots, descriptive statistics, etc.), as hopefully illustrated by this paper or by Insa-Cabrera et al. (2011), and not just a plot of the best results achieved, as shown by Legg and Veness (2013). Since it requires searching a large configuration space, the weighted statistics approach proposed in Section 4.2.7 should be considered as well as the effectiveness of the testing process. For the AIQ test this was discussed in Sections 4.2.5 and 4.2.6. In the AIT proposal, Hernández-Orallo and Dowe (2010) also discuss this issue extensively.

The 'what for' question is basically a question of motivation. Obviously, the primary use of such a method is to evaluate AI or AGI systems, however as Vadinský (2015) suggests, generalizations

from the results of particular systems can be made to the level of a certain AI or AGI paradigm. These paradigms can then be compared, and even some fundamental questions of AI could be (at least partially) answered and grounded in experimental results. Furthermore, the evaluation of an AI or AGI system need not only be external, it could be internalized to some form of meta-learning that would enable the system to fine-tune its parameters to the different needs of specific tasks. Based on knowledge of the performance of its configurations in a general test, especially if broken down to a more fine-grained level as proposed in Section 4.2.4, this meta-learning could be guided by prioritizing the generally more promising configurations.

5. Conclusion and Future Work

This paper attempted to replicate the results Legg and Veness (2013, 2011) presented along with the transformation of the Universal Intelligence definition into a practicable AIQ test. Three experiments were conducted: the default setting, and variation of both the action and observation space settings. The replication of the original results has only been partially successful:

- The score of *freq*, Q_0, Q_λ, and HLQ_λ in the default setting matched the original results, while for *MC-AIXI* it was higher for the short episodes and lower for the long episodes.
- Overall results when varying the observation space setting were "qualitatively the same" as the default setting as reported by Legg and Veness.
- Overall results when varying the action space setting however were not "qualitatively the same", contrary to the results of Legg and Veness.

In cases when the exact configuration of agents and all the test parameters were known, *the replicated results match the original results rather well, suggesting that the test is reliable*. Since insufficient details are given by Legg and Veness about the variation of the action and observation space settings, the reliability of the test in these cases cannot be assessed. However, as all of the replication experiment settings used the same agent configurations, the sensitivity of the test to the changes in the BF machine settings can be evaluated. The discrepancy in the conclusions of the two experiments is likely only apparent, since increasing the size of the observation space does not in fact increase the number of produced observations in all environments, causing the resulting increase in complexity to be less than expected. Therefore, the results from the experiment varying the size of action space dominate, and it can be concluded that *the AIQ test is sensitive to changes in the BF reference machine settings* contrary to the original claim of Legg and Veness. An additional contribution of the paper lies in a precise specification of the conducted experiment settings and the achieved results. This is in contrast to Legg and Veness who are rather vague in this regard.

Since Legg and Veness (2013, 2011) do not state the precise configurations of *MC-AIXI* used in the original experiments, this case should be interpreted with caution. On inquiry, Joel Veness helped to clarify what settings were used, but could no longer provide exact values. This is most likely the reason why the replicated results do not match. Concerning the performance of *MC-AIXI* with respect to the episode length, Legg and Veness (2013) note that "Our initial attempts at modifying MC-AIXI to be similarly high scoring on shorter runs failed." As the conducted parameter sweep shows, restricting the world model by lowering the *number of Monte Carlo simulations*, *context tree depth*, and *search horizon* increases its performance for short episodes. Furthermore, since a total of 840 distinct configurations were tested, the results were analyzed for

any impact of modifying parameter values on the performance of an agent. Notably, the following observations concerning the performance of *MC-AIXI* configurations in the default experiment were made:

- The overall performance of configurations using exponentially decaying exploration is substantially better than those featured constant exploration values.
- Configurations with a *search horizon* of 1 perform rather poorly. When its value is increased to 2 or 3, the results improve considerably, however, further increasing the parameter value has a slightly negative effect (although this effect diminishes with increasing episode length).
- Configurations with high *exploration decay* values have a rather lower spread of their results at long episodes than with lower *exploration decay*.

A detailed structure of the influence of parameter values was discovered using data mining techniques. This way several highly (as well as poorly) performing configurations were identified. This explorative analysis also contributes by illustrating a possible use for any suitable AGI evaluation method.

The final contribution of the paper rests in assessing the AIQ test regarding its suitability as a general AI evaluation method. Based on the literature overview and the conducted experiments, several limits of the AIQ test were noted and some ways to amend them were suggested. While some of the proposals were only discussed, some others were implemented or demonstrated:

- There is a dependence of the AIQ test on the chosen reference machine (and its parameters) which could be reduced by enforcing a limit on the minimum length of environment programs as suggested by Hibbard (2009). The number of reference machine instructions was proposed as a possible value of such a limit for further investigation. This proposal was implemented in the extended test and the properties of the resulting new program samples were demonstrated. An alternative solutional approach of Hernández-Orallo (2015, 2017) to reducing the dependence on the reference machine was also discussed.
- Although the AIQ test is guarded against some non-discriminative environments, others still remain. However, the proposal of Hernández-Orallo and Dowe (2010) to use only reward-sensitive environments seems too strong (as it also excludes those containing "heaven" and "hell" subenvironments). Instead, it was proposed to extend the capabilities of the AIQ test to exclude particular types of problematic environment programs.
- The AIQ score contains some aspects of intelligence only implicitly, while others it does not contain at all. There are proposals e.g. by Goertzel (2010) or by Hernández-Orallo and Dowe (2010) to integrate those into the score. While this could be helpful, it would also add further dimensions of uncertainty when interpreting the results. Thus, it was proposed that the overall score be supplemented by also providing suitable scores for particular aspects of intelligence.
- Since the AIQ test reveals rather limited differences among the tested agents, there is a concern that the test setting may be too simple. Several approaches were proposed and demonstrated which would help to answer this concern, if conducted in sufficient depth.
- The test format highlights the computational requirements of an agent. Among others, a procedure was proposed to achieve sufficiently converged estimates of the AIQ score by

progressively increasing the demanding parameters of the AIQ test only if the chosen stopping condition had not yet been met. One of the proposed methods was implemented, and its benefits were demonstrated. This method both saves some resources and also facilitates the implementation of other proposed methods.

- Since there are differences in the dynamics of the AIQ score convergence process, some characteristics were proposed to describe it. These could also serve as suitable stopping criteria in the previously-mentioned testing procedure.

- Noting the differences in the results of agent configurations, a thorough parameter sweep is needed to research its capacities, the results of which should be reported statistically. In case of prior knowledge about the agent resulting in a non-random sample of its configurations, a method to weight such statistics was proposed and demonstrated using the example of *MC-AIXI*.

While the methods to amend the limits of the AIQ test were proposed and discussed, and some were even implemented, there remain many opportunities for future work involving both the implementation and testing of the remaining proposals. The following areas are of particular interest for future work since they could benefit from the precisely stated results of the conducted experiments:

- A detailed analysis using other reference machines as well as the newly implemented minimal program length restriction would bring empirical light into the debate on the extent of reference machine dependence of the AIQ test.

- An in-depth comparison with the Anytime Intelligence Test prototype would further help with AIQ test evaluation, as well as with investigating the possible simplicity of the test setting.

- A better understanding of the AIQ test inner workings is still needed. The preliminary results of semantic analysis of the environment programs seem promising, however the analysis is still incomplete and the resulting program classes should be investigated in more depth.

Acknowledgments

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

Further thanks are due to anonymous reviewers of AGI Conference 2016 where an earlier version of this paper was submitted, to Joel Veness who answered author's questions regarding the settings of *MC-AIXI* in the original paper, to Ondřej Havlíček who helped the author to better understand the necessary methods of statistical analysis, and to anonymous reviewers of JAGI for their inspiring comments.

References

Besold, T.; Hernández-Orallo, J.; and Schmid, U. 2015. Can Machine Intelligence be Measured in the Same Way as Human intelligence? *KI – Künstliche Intelligenz* 29(3):291–297.

Breiman, L.; Friedman, J. H.; Olsen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees*. Belmont: Thomson Wadsworth.

Bringsjord, S., and Schimanski, B. 2003. What Is Artificial Intelligence? Psychometric AI as an Answer. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, 887–893.

de Mey, M. 1992. *The Cognitive Paradigm*. Chicago and London: University of Chicago Press.

Dennett, D. C. 1991. *Consciousness Explained*. London: Penguin Books.

Descartes, R. 1637. *A Discourse on Method*. Oxford: Oxford University Press.

Dowe, D. L., and Hájek, A. R. 1998. A Non-Behavioural, Computational Extension to the Turing Test. In *Proceedings of International Conference on Computational Intelligence & Multimedia Applications (ICCIMA'98), Gippsland, Australia*, 101–106.

Goertzel, B. 2010. Toward a Formal Characterization of Real-World General Intelligence. In Baum, E.; Hutter, M.; and Kitzelmann, E., eds., *Proceedings of the 3rd Conference on Artificial General Intelligence, AGI 2010*, 19–24. Amsterdam-Beijing-Paris: Atlantis Press.

Goertzel, B. 2014. Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence* 5(1):1–48.

Harnad, S. 1991. Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem. *Minds and Machines* 1(1):43–54.

Hernández-Orallo, J., and Dowe, D. L. 2010. Measuring Universal Intelligence: Towards an Anytime Intelligence Test. *Artificial Intelligence* 174(18):1508–1539.

Hernandez-Orallo, J. 2000. Beyond the Turing Test. *Journal of Logic, Language and Information* 9(4):447–466.

Hernández-Orallo, J. 2010. A (hopefully) Unbiased Universal Environment Class for Measuring Intelligence of Biological and Artificial Systems. In Baum, E.; Hutter, M.; and Kitzelmann, E., eds., *Proceedings of the 3rd Conference on Artificial General Intelligence, AGI 2010*, 182–183. Amsterdam-Beijing-Paris: Atlantis Press.

Hernández-Orallo, J. 2015. C-Tests Revisited: Back and Forth with Complexity. In Bieger, J.; Goertzel, B.; and Potapov, A., eds., *Proceedings of the 8th Conference on Artificial General Intelligence, AGI 2015*, volume 9205 of *Lecture notes in artificial intelligence*, 272–282. Berlin: Springer.

Hernández-Orallo, J. 2017. *The Measure of All Minds*. Cambridge: Cambridge University Press.

Hibbard, B. 2009. Bias and No Free Lunch in Formal Measures of Intelligence. *Journal of Artificial General Intelligence* 1(1):54–61.

Hothorn, T.; Hornik, K.; and Zeileis, A. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 3(15):651–674.

Hutter, M., and Legg, S. 2007. Temporal Difference Updating without a Learning Rate. In Platt, J. C.; Koller, D.; Singer, Y.; and Roweis, S. T., eds., *Advances in Neural Information Processing Systems 20*, 705–712. Curran Associates, Inc.

Insa-Cabrera, J.; Dowe, D. L.; España-Cubillo, S.; Hernández-Lloreda, M. V.; and Hernández-Orallo, J. 2011. Comparing Humans and AI Agents. In Schmidhuber, J.; Thórisson, K. R.; and Looks, M., eds., *Proceedings of the 4th Conference on Artificial General Intelligence, AGI 2011*, volume 6830 of *Lecture notes in artificial intelligence*, 122–132. Berlin: Springer.

Legg, S., and Hutter, M. 2007a. A Collection of Definitions of Intelligence. In Goertzel, B., and Wang, P., eds., *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, volume 157 of *Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press. 17–24.

Legg, S., and Hutter, M. 2007b. Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines* 17(4):391–444.

Legg, S., and Veness, J. 2011. AIQ: Algorithmic Intelligence Quotient [source codes]. `https://github.com/mathemajician/AIQ`. Accessed: 2017-06-26.

Legg, S., and Veness, J. 2013. An Approximation of the Universal Intelligence Measure. In Dowe, D. L., ed., *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*, volume 7070 of *Lecture Notes in Computer Science*. Berlin: Springer. 236–249.

Müller, U. 1993. dev/lang/brainfuck-2.lha in Aminet. `http://aminet.net/package.php?package=dev/lang/brainfuck-2.lha`. Accessed: 2017-06-26.

Schweizer, P. 2012. The Externalist Foundations of a Truly Total Turing Test. *Minds and Machines* 22(3):191–212.

Searle, J. R. 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences* 3(3):417–457.

Sun, R. 2007. The Importance of Cognitive Architectures: An Analysis Based on CLARION. *Journal of Experimental & Theoretical Artificial Intelligence* 19(2):159–193.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59(236):433–460.

Vadinský, O. 2015. Towards an Artificially Intelligent System: Possibilities of General Evaluation of Hybrid Paradigm. In Besold, T. R.; Lamb, L. C.; Icard, T.; and Miikkulainen, R., eds., *Proceedings of the 10th International Workshop on Neural-Symbolic Learning and Reasoning NeSy'15*, 23–29. Buenos Aires: IJCAI.

Veness, J.; Ng, K. S.; Hutter, M.; Uther, W.; and Silver, D. 2011. A Monte Carlo AIXI Approximation. *Journal of Artificial Intelligence Research* 40(1):95–142.

Watkins, C. 1989. *Learning from Delayed Rewards*. Ph.D. Dissertation, Kings College, Cambridge, England.

Table 7: Estimated AIQ scores with a 0.95 confidence interval for all tested agent configurations. Agent's extremes in italics, overall extremes in bold. Agent's parameters explained in Section 2.6 (BF 5 Reference Machine).

Agent	**Configuration**		**AIQ Scores with Confidence Intervals for Episode Length**									
	#	*Parameters*	*1,000*		*3,000*		*10,000*		*30,000*		*100,000*	
freq	1	(0.03)	*39.5*	*± 0.5*	42.1	± 0.5	***43.9***	***± 0.5***	***45.9***	***± 0.5***	47.7	± 0.4
freq	2	(0.05)	*40.1*	*± 0.5*	42.3	± 0.5	44.5	± 0.5	46.6	± 0.5	*48.1*	*± 0.4*
freq	3	(0.07)	*40.1*	*± 0.5*	*42.8*	*± 0.5*	44.8	± 0.4	*46.9*	*± 0.4*	48.0	± 0.4
freq	4	(0.09)	39.6	± 0.4	42.5	± 0.4	*44.9*	*± 0.4*	46.6	± 0.4	47.4	± 0.4
freq	5	(0.11)	39.7	± 0.4	*41.7*	*± 0.4*	*44.9*	*± 0.4*	***45.9***	***± 0.4***	***46.8***	***± 0.4***
Q_0	1	(0,0,0.5,0.04,0.6)	*41.8*	*± 0.4*	45.5	± 0.4	46.6	± 0.4	*46.9*	*± 0.4*	*47.0*	*± 0.4*
Q_0	2	(0,0,0.5,0.03,0.7)	41.6	± 0.4	*46.1*	*± 0.4*	48.0	± 0.4	49.1	± 0.4	48.9	± 0.4
Q_0	3	(0,0,0.5,0.02,0.8)	40.4	± 0.5	46.0	± 0.5	*49.0*	*± 0.4*	50.7	± 0.4	50.9	± 0.4
Q_0	4	(0,0,0.5,0.01,0.9)	37.9	± 0.5	42.5	± 0.5	48.5	± 0.5	*51.4*	*± 0.4*	52.7	± 0.4
Q_0	5	(0,0,0.5,0.005,0.95)	*37.1*	*± 0.5*	***39.2***	***± 0.5***	*44.4*	*± 0.5*	49.7	± 0.4	*53.0*	*± 0.4*
Q_λ	1	(0,0.5,0.5,0.04,0.6)	*44.0*	*± 0.4*	47.8	± 0.4	48.5	± 0.4	*48.9*	*± 0.4*	*49.1*	*± 0.4*
Q_λ	2	(0,0.5,0.5,0.03,0.6)	*44.0*	*± 0.4*	*48.1*	*± 0.4*	49.4	± 0.4	50.1	± 0.4	50.1	± 0.4
Q_λ	3	(0,0.5,0.5,0.02,0.8)	42.5	± 0.5	47.6	± 0.5	*50.6*	*± 0.4*	51.6	± 0.4	52.2	± 0.4
Q_λ	4	(0,0.5,0.5,0.01,0.9)	40.0	± 0.5	43.9	± 0.5	49.3	± 0.5	*52.3*	*± 0.4*	53.4	± 0.4
Q_λ	5	(0,0.5,0.5,0.005,0.95)	*39.0*	*± 0.5*	*40.6*	*± 0.5*	*46.0*	*± 0.5*	51.1	± 0.5	*53.8*	*± 0.4*
HLQ_λ	1	(0,0,0.99,0.02,0.7)	46.5	± 0.5	***51.9***	***± 0.5***	***53.8***	***± 0.5***	***54.7***	***± 0.5***	55.0	± 0.4
HLQ_λ	2	(0,0,0.95,0.04,0.7)	***48.6***	***± 0.5***	50.8	± 0.4	51.9	± 0.4	*51.9*	*± 0.4*	*52.0*	*± 0.4*
HLQ_λ	3	(0,0,0.99,0.04,0.6)	48.3	± 0.5	51.0	± 0.4	53.3	± 0.4	53.5	± 0.4	53.4	± 0.4
HLQ_λ	4	(0,0,0.995,0.01,0.8)	42.8	± 0.5	49.3	± 0.5	53.7	± 0.5	54.5	± 0.4	***55.2***	***± 0.4***
HLQ_λ	5	(0,0,0.995,0.005,0.9)	*40.3*	*± 0.5*	*46.2*	*± 0.5*	*51.5*	*± 0.5*	54.6	± 0.5	55.0	± 0.5
MC-AIXI	1	(50,8,3,0.85,0.3)	42.1	± 1.2	*50.2*	*± 1.3*	***54.0***	***± 1.4***	53.8	± 1.4	54.3	± 1.4
MC-AIXI	2	(50,16,2,0.95,0.9)	*44.5*	*± 1.1*	48.6	± 1.1	51.2	± 1.2	51.5	± 1.3	51.8	± 1.2
MC-AIXI	3	(100,8,3,0.9,0.6)	38.8	± 1.1	48.5	± 1.2	53.6	± 1.4	***54.4***	***± 1.4***	***55.3***	***± 1.4***
MC-AIXI	4	(100,16,4,0.85,0.9)	40.3	± 1.1	48.3	± 1.2	51.8	± 1.3	52.9	± 1.3	53.5	± 1.3
MC-AIXI	5	(100,32,3,0.05,1.0)	***36.1***	***± 1.1***	*42.1*	*± 1.1*	*46.8*	*± 1.2*	*48.8*	*± 1.3*	*50.5*	*± 1.3*

Appendix A. Tables and Plots of Results from All Experiments

The detailed results of the default experiment described in Section 3.1 are given by Table 7. All tested configurations of *freq*, Q_0, Q_λ, and HLQ_λ are shown, while only selected configurations of *MC-AIXI* are listed to illustrate its performance. Table 8 gives descriptive statistics of *MC-AIXI* results computed from all tested configurations. Full results are available from: `https://github.com/xvado00/RATR/archive/v1.0.zip`.

The detailed results of the experiment that varies the action space as described in Section 3.2 are given by Tables 9, 10, and 11.

The detailed results of the experiment that varies the observation space as described in Section 3.3 are given by Tables 12, and 13.

Figure 9 shows the best achieved AIQ score estimates with a margin of error corresponding to a 0.95 confidence interval for each agent after the tested number of interactions in all experiments.

Table 8: Statistics of *MC-AIXI* AIQ scores for Episode Length (BF 5 reference machine)

Measure	**AIQ Score Statistics for Episode Length**				
	1,000	*3,000*	*10,000*	*30,000*	*100,000*
Mean	35.07	42.21	45.68	46.84	47.52
Standard Error	0.15	0.16	0.19	0.20	0.20
Minimum	21.9	25.1	24.5	23.7	23.4
First Quartile	32.2	39.9	42.875	44	44.4
Median	35.2	43.35	47.4	48.65	49.4
Third Quartile	38.4	45.2	49.4	50.825	51.7
Maximum	44.5	50.5	54	54.8	55.3
Standard Deviation	4.37	4.61	5.37	5.74	5.89
Kurtosis	−0.31	1.09	1.45	1.81	1.94
Skewness	−0.37	−1.06	−1.27	−1.36	−1.38

Table 9: Estimated AIQ scores with a 0.95 confidence interval for all tested agent configurations. Agent's extremes in italics, overall extremes in bold. Agent's parameters explained in Section 2.6 (BF 2 Reference Machine).

Agent	**Configuration**		**AIQ Scores with Confidence Intervals for Episode Length**									
	#	*Parameters*	*1,000*		*3,000*		*10,000*		*30,000*		*100,000*	
freq	1	(0.03)	***28.2***	***±0.6***	***33.6***	***±0.5***	***41.4***	***±0.5***	***46.7***	***±0.5***	50.3	±0.4
freq	2	(0.05)	31.2	±0.5	39.2	±0.5	45.3	±0.5	48.3	±0.5	*50.4*	*±0.4*
freq	3	(0.07)	34.6	±0.5	42.2	±0.5	46.5	±0.5	*49.4*	*±0.4*	49.9	±0.4
freq	4	(0.09)	36.6	±0.5	42.2	±0.5	*46.9*	*±0.5*	48.5	±0.4	49.6	±0.4
freq	5	(0.11)	*38.2*	*±0.5*	*43.2*	*±0.5*	45.9	±0.4	48.0	±0.4	*48.6*	*±0.4*
Q_0	1	(0,0,0.5,0.04,0.6)	*50.5*	*±0.4*	*51.0*	*±0.4*	*51.1*	*±0.4*	*51.0*	*±0.4*	*51.0*	*±0.4*
Q_0	2	(0,0,0.5,0.03,0.7)	51.6	±0.4	52.3	±0.4	52.4	±0.4	52.4	±0.4	52.2	±0.4
Q_0	3	(0,0,0.5,0.02,0.8)	52.8	±0.4	53.6	±0.4	53.8	±0.4	53.9	±0.4	53.8	±0.4
Q_0	4	(0,0,0.5,0.01,0.9)	53.5	±0.4	54.6	±0.4	55.3	±0.4	55.3	±0.4	55.2	±0.4
Q_0	5	(0,0,0.5,0.005,0.95)	*53.6*	*±0.5*	*55.0*	*±0.4*	*55.8*	*±0.4*	*56.1*	*±0.4*	*56.0*	*±0.4*
Q_λ	1	(0,0.5,0.5,0.04,0.6)	*52.4*	*±0.4*	*52.9*	*±0.4*	*53.0*	*±0.4*	*53.0*	*±0.4*	*52.9*	*±0.4*
Q_λ	2	(0,0.5,0.5,0.03,0.6)	53.6	±0.4	53.9	±0.4	54.0	±0.4	54.2	±0.4	53.8	±0.4
Q_λ	3	(0,0.5,0.5,0.02,0.8)	55.0	±0.4	55.4	±0.4	55.6	±0.4	55.7	±0.4	55.3	±0.4
Q_λ	4	(0,0.5,0.5,0.01,0.9)	55.6	±0.4	56.7	±0.4	57.0	±0.4	57.1	±0.4	56.9	±0.4
Q_λ	5	(0,0.5,0.5,0.005,0.95)	*55.6*	*±0.5*	*57.0*	*±0.4*	***57.7***	***±0.4***	***57.9***	***±0.4***	***57.8***	***±0.4***
HLQ_λ	1	(0,0,0.99,0.02,0.7)	54.9	±0.4	*55.2*	*±0.4*	*55.3*	*±0.4*	*55.3*	*±0.4*	*55.0*	*±0.4*
HLQ_λ	2	(0,0,0.95,0.04,0.7)	*54.7*	*±0.4*	55.3	±0.4	55.6	±0.4	55.6	±0.4	55.3	±0.4
HLQ_λ	3	(0,0,0.99,0.04,0.6)	55.9	±0.4	56.7	±0.4	56.8	±0.4	56.8	±0.4	56.6	±0.4
HLQ_λ	4	(0,0,0.995,0.01,0.8)	56.5	±0.4	57.3	±0.4	57.5	±0.4	57.5	±0.4	57.3	±0.4
HLQ_λ	5	(0,0,0.995,0.005,0.9)	***56.9***	***±0.4***	***57.5***	***±0.4***	***57.8***	***±0.4***	***57.9***	***±0.4***	***57.5***	***±0.4***
MC-AIXI	1	(50,8,3,0.85,0.3)	45.0	±1.6	*45.4*	*±1.7*	47.2	±1.8	***46.8***	***±1.7***	***46.2***	***±1.7***
MC-AIXI	2	(50,16,2,0.95,0.9)	*54.1*	*±1.2*	55.6	±1.3	56.1	±1.3	56.0	±1.3	55.4	±1.3
MC-AIXI	3	(100,8,3,0.9,0.6)	*44.0*	*±1.4*	45.5	±1.5	*45.5*	*±1.5*	45.2	±1.5	45.2	±1.4
MC-AIXI	4	(100,16,4,0.85,0.9)	53.8	±1.2	*55.9*	*±1.3*	***57.4***	***±1.3***	***57.6***	***±1.3***	***57.1***	***±1.3***
MC-AIXI	5	(100,32,3,0.05,1.0)	51.9	±1.2	53.7	±1.2	54.7	±1.2	54.4	±1.2	54.3	±1.2

Table 10: Estimated AIQ scores with a 0.95 confidence interval for all tested agent configurations. Agent's extremes in italics, overall extremes in bold. Agent's parameters explained in Section 2.6 (BF 10 Reference Machine).

Agent	**Configuration**		**AIQ Scores with Confidence Intervals for Episode Length**									
	#	*Parameters*	*1,000*		*3,000*		*10,000*		*30,000*		*100,000*	
freq	1	(0.03)	*37.4*	*± 0.5*	*40.9*	*± 0.5*	*44.2*	*± 0.5*	*47.1*	*± 0.5*	49.6	± 0.5
freq	2	(0.05)	38.5	± 0.5	42.0	± 0.4	45.7	± 0.4	*48.4*	*± 0.5*	*50.1*	*± 0.5*
freq	3	(0.07)	38.9	± 0.4	42.8	± 0.4	46.0	± 0.4	48.3	± 0.5	49.7	± 0.5
freq	4	(0.09)	38.9	± 0.4	42.9	± 0.4	*46.4*	*± 0.4*	48.0	± 0.4	49.4	± 0.5
freq	5	(0.11)	*39.0*	*± 0.4*	*43.1*	*± 0.4*	45.9	± 0.4	47.4	± 0.4	*48.4*	*± 0.4*
Q_0	1	(0,0,0.5,0.04,0.6)	*35.9*	*± 0.4*	*40.9*	*± 0.4*	44.4	± 0.4	45.9	± 0.4	*46.5*	*± 0.4*
Q_0	2	(0,0,0.5,0.03,0.7)	35.4	± 0.4	*40.9*	*± 0.4*	*45.4*	*± 0.4*	47.5	± 0.4	48.2	± 0.4
Q_0	3	(0,0,0.5,0.02,0.8)	33.9	± 0.4	39.6	± 0.4	*45.4*	*± 0.4*	*48.3*	*± 0.5*	49.5	± 0.5
Q_0	4	(0,0,0.5,0.01,0.9)	30.4	± 0.4	35.2	± 0.4	42.1	± 0.5	47.9	± 0.5	*51.1*	*± 0.5*
Q_0	5	(0,0,0.5,0.005,0.95)	*29.0*	*± 0.4*	***30.6***	***± 0.4***	***36.7***	***± 0.4***	***44.2***	***± 0.5***	49.8	± 0.5
Q_λ	1	(0,0.5,0.5,0.04,0.6)	*38.8*	*± 0.4*	43.7	± 0.4	47.3	± 0.4	48.6	± 0.4	*49.3*	*± 0.4*
Q_λ	2	(0,0.5,0.5,0.03,0.6)	38.6	± 0.4	*44.0*	*± 0.4*	47.6	± 0.4	49.5	± 0.4	50.2	± 0.4
Q_λ	3	(0,0.5,0.5,0.02,0.8)	36.5	± 0.4	42.4	± 0.4	*47.7*	*± 0.5*	*50.9*	*± 0.5*	51.9	± 0.5
Q_λ	4	(0,0.5,0.5,0.01,0.9)	33.5	± 0.4	37.6	± 0.4	44.5	± 0.5	49.8	± 0.5	*52.4*	*± 0.5*
Q_λ	5	(0,0.5,0.5,0.005,0.95)	*31.9*	*± 0.4*	*34.0*	*± 0.4*	*38.8*	*± 0.5*	*45.7*	*± 0.5*	51.0	± 0.5
HLQ_λ	1	(0,0,0.99,0.02,0.7)	***44.5***	***± 0.4***	48.9	± 0.4	51.2	± 0.5	*52.1*	*± 0.4*	*52.5*	*± 0.4*
HLQ_λ	2	(0,0,0.95,0.04,0.7)	44.3	± 0.4	***49.6***	***± 0.5***	52.6	± 0.5	53.9	± 0.5	54.0	± 0.5
HLQ_λ	3	(0,0,0.99,0.04,0.6)	42.2	± 0.5	48.9	± 0.5	***53.0***	***± 0.5***	***54.8***	***± 0.5***	55.3	± 0.5
HLQ_λ	4	(0,0,0.995,0.01,0.8)	37.9	± 0.5	45.6	± 0.5	52.1	± 0.5	***54.8***	***± 0.5***	***55.9***	***± 0.5***
HLQ_λ	5	(0,0,0.995,0.005,0.9)	*34.3*	*± 0.5*	*41.7*	*± 0.5*	*49.1*	*± 0.5*	53.5	± 0.5	55.4	± 0.5
MC-AIXI	1	(50,8,3,0.85,0.3)	33.6	± 1.0	42.6	± 1.3	*47.6*	*± 1.4*	49.4	± 1.4	48.2	± 1.3
MC-AIXI	2	(50,16,2,0.95,0.9)	*35.0*	*± 1.1*	*43.1*	*± 1.3*	47.5	± 1.5	49.2	± 1.5	49.4	± 1.6
MC-AIXI	3	(100,8,3,0.9,0.6)	28.0	± 1.1	39.6	± 1.3	47.1	± 1.4	49.6	± 1.5	49.6	± 1.5
MC-AIXI	4	(100,16,4,0.85,0.9)	26.4	± 1.0	36.5	± 1.3	47.0	± 1.3	*51.8*	*± 1.5*	*51.4*	*± 1.6*
MC-AIXI	5	(100,32,3,0.05,1.0)	***27.2***	***± 0.9***	*33.7*	*± 1.0*	*41.4*	*± 1.2*	***44.2***	***± 1.2***	***45.8***	***± 1.3***

Table 11: Estimated AIQ scores with a 0.95 confidence interval for all tested agent configurations. Agent's extremes in italics, overall extremes in bold. Agent's parameters explained in Section 2.6 (BF 20 Reference Machine).

Agent	**Configuration**		**AIQ Scores with Confidence Intervals for Episode Length**									
	#	*Parameters*	*1,000*		*3,000*		*10,000*		*30,000*		*100,000*	
freq	1	(0.03)	*35.7*	*± 0.4*	*38.4*	*± 0.4*	*41.5*	*± 0.4*	*45.5*	*± 0.4*	*48.1*	*± 0.5*
freq	2	(0.05)	36.1	± 0.4	39.1	± 0.4	43.7	± 0.4	46.8	± 0.4	48.8	± 0.5
freq	3	(0.07)	37.0	± 0.4	40.4	± 0.4	44.5	± 0.4	*47.2*	*± 0.4*	*48.9*	*± 0.4*
freq	4	(0.09)	37.0	± 0.4	*41.2*	*± 0.4*	44.8	± 0.4	47.1	± 0.4	48.6	± 0.4
freq	5	(0.11)	*37.2*	*± 0.4*	*41.2*	*± 0.4*	*44.9*	*± 0.4*	46.5	± 0.4	*48.1*	*± 0.4*
Q_0	1	(0,0,0.5,0.04,0.6)	*31.9*	*± 0.4*	*36.7*	*± 0.4*	40.9	± 0.4	43.3	± 0.4	*44.5*	*± 0.4*
Q_0	2	(0,0,0.5,0.03,0.7)	31.2	± 0.4	36.3	± 0.4	*41.1*	*± 0.4*	44.3	± 0.4	45.7	± 0.4
Q_0	3	(0,0,0.5,0.02,0.8)	30.0	± 0.4	34.8	± 0.4	40.5	± 0.4	*44.6*	*± 0.4*	46.7	± 0.4
Q_0	4	(0,0,0.5,0.01,0.9)	27.5	± 0.4	30.6	± 0.4	36.9	± 0.4	42.6	± 0.4	*46.8*	*± 0.5*
Q_0	5	(0,0,0.5,0.005,0.95)	*26.2*	*± 0.4*	*27.8*	*± 0.4*	*31.3*	*± 0.4*	*37.9*	*± 0.4*	44.7	± 0.5
Q_λ	1	(0,0.5,0.5,0.04,0.6)	*35.2*	*± 0.4*	*39.7*	*± 0.4*	43.9	± 0.4	46.1	± 0.4	47.0	± 0.4
Q_λ	2	(0,0.5,0.5,0.03,0.6)	34.8	± 0.4	39.6	± 0.4	*44.1*	*± 0.4*	46.5	± 0.4	47.7	± 0.4
Q_λ	3	(0,0.5,0.5,0.02,0.8)	32.6	± 0.4	37.3	± 0.4	43.3	± 0.4	*47.1*	*± 0.4*	*49.2*	*± 0.4*
Q_λ	4	(0,0.5,0.5,0.01,0.9)	30.2	± 0.4	32.8	± 0.4	38.7	± 0.4	44.8	± 0.5	48.9	± 0.5
Q_λ	5	(0,0.5,0.5,0.005,0.95)	*29.6*	*± 0.4*	*30.7*	*± 0.4*	*33.8*	*± 0.4*	*39.9*	*± 0.5*	*46.2*	*± 0.5*
HLQ_λ	1	(0,0,0.99,0.02,0.7)	***40.5***	***± 0.4***	45.2	± 0.4	48.1	± 0.4	*49.5*	*± 0.4*	*50.1*	*± 0.4*
HLQ_λ	2	(0,0,0.95,0.04,0.7)	40.1	± 0.4	***46.1***	***± 0.4***	49.5	± 0.4	51.1	± 0.4	51.6	± 0.5
HLQ_λ	3	(0,0,0.99,0.04,0.6)	38.6	± 0.4	45.0	± 0.5	***49.7***	***± 0.5***	***52.0***	***± 0.5***	53.1	± 0.5
HLQ_λ	4	(0,0,0.995,0.01,0.8)	34.5	± 0.4	41.9	± 0.5	48.4	± 0.5	***52.0***	***± 0.5***	***53.6***	***± 0.5***
HLQ_λ	5	(0,0,0.995,0.005,0.9)	*30.8*	*± 0.4*	*38.0*	*± 0.4*	*45.5*	*± 0.5*	50.3	± 0.5	52.7	± 0.5
MC-AIXI	1	(50,8,3,0.85,0.3)	11.7	± 0.7	***14.9***	***± 0.9***	***18.6***	***± 1.2***	***22.6***	***± 1.3***	***26.4***	***± 1.4***
MC-AIXI	2	(50,16,2,0.95,0.9)	*27.2*	*± 0.9*	*34.9*	*± 1.0*	*38.9*	*± 1.1*	*42.2*	*± 1.3*	42.1	± 1.3
MC-AIXI	3	(100,8,3,0.9,0.6)	***10.7***	***± 0.8***	16.8	± 1.0	22.1	± 1.2	27.1	± 1.5	29.0	± 1.5
MC-AIXI	4	(100,16,4,0.85,0.9)	21.0	± 0.6	28.6	± 0.8	34.9	± 0.9	39.5	± 1.1	*42.3*	*± 1.3*
MC-AIXI	5	(100,32,3,0.05,1.0)	21.4	± 0.7	25.1	± 0.7	28.9	± 0.7	34.0	± 0.8	37.4	± 1.0

Table 12: Estimated AIQ scores with a 0.95 confidence interval for all tested agent configurations. Agent's extremes in italics, overall extremes in bold. Agent's parameters explained in Section 2.6 (BF 5,2 Reference Machine).

Agent	**Configuration**		**AIQ Scores with Confidence Intervals for Episode Length**									
	#	*Parameters*	*1,000*		*3,000*		*10,000*		*30,000*		*100,000*	
freq	1	(0.03)	*38.6*	± *0.5*	42.1	± 0.5	***43.9***	***± 0.5***	***45.6***	***± 0.5***	47.7	± 0.5
freq	2	(0.05)	*39.7*	± *0.5*	*42.0*	± *0.5*	44.3	± 0.5	46.4	± 0.4	*48.6*	± *0.4*
freq	3	(0.07)	39.5	± 0.5	*42.7*	± *0.5*	*44.8*	± *0.5*	*46.7*	± *0.5*	48.0	± 0.4
freq	4	(0.09)	39.3	± 0.4	42.5	± 0.5	44.6	± 0.4	46.5	± 0.5	47.7	± 0.4
freq	5	(0.11)	39.4	± 0.4	42.1	± 0.4	44.1	± 0.4	46.1	± 0.4	*47.1*	± *0.4*
Q_0	1	(0,0,0.5,0.04,0.6)	42.3	± 0.4	45.1	± 0.4	*45.8*	± *0.4*	*46.0*	± *0.4*	***45.9***	***± 0.4***
Q_0	2	(0,0,0.5,0.03,0.7)	42.4	± 0.4	46.5	± 0.4	47.7	± 0.4	48.0	± 0.4	48.2	± 0.4
Q_0	3	(0,0,0.5,0.02,0.8)	41.9	± 0.5	46.9	± 0.5	49.3	± 0.4	50.2	± 0.4	50.3	± 0.4
Q_0	4	(0,0,0.5,0.01,0.9)	39.4	± 0.5	43.9	± 0.5	49.3	± 0.5	51.9	± 0.5	52.7	± 0.4
Q_0	5	(0,0,0.5,0.005,0.95)	*38.4*	± *0.5*	***40.2***	***± 0.5***	46.3	± 0.5	50.8	± 0.5	53.7	± 0.5
Q_λ	1	(0,0.5,0.5,0.04,0.6)	44.5	± 0.4	47.2	± 0.4	48.5	± 0.4	*48.0*	± *0.4*	*48.8*	± *0.4*
Q_λ	2	(0,0.5,0.5,0.03,0.6)	*44.7*	± *0.4*	47.8	± 0.4	48.9	± 0.4	49.3	± 0.4	49.3	± 0.4
Q_λ	3	(0,0.5,0.5,0.02,0.8)	43.8	± 0.5	*47.9*	± *0.4*	50.6	± 0.4	51.8	± 0.4	51.7	± 0.4
Q_λ	4	(0,0.5,0.5,0.01,0.9)	41.3	± 0.5	45.4	± 0.5	*51.1*	± *0.5*	*53.0*	± *0.5*	53.4	± 0.4
Q_λ	5	(0,0.5,0.5,0.005,0.95)	*39.9*	± *0.5*	*43.0*	± *0.5*	*47.8*	± *0.5*	51.6	± 0.5	*53.8*	± *0.5*
HLQ_λ	1	(0,0,0.99,0.02,0.7)	***48.8***	***± 0.5***	51.1	± 0.4	*51.7*	± *0.4*	*51.7*	± *0.4*	*51.8*	± *0.4*
HLQ_λ	2	(0,0,0.95,0.04,0.7)	48.5	± 0.5	***52.0***	***± 0.5***	53.1	± 0.4	53.3	± 0.4	53.4	± 0.4
HLQ_λ	3	(0,0,0.99,0.04,0.6)	47.3	± 0.5	***52.0***	***± 0.5***	***53.9***	***± 0.5***	54.6	± 0.4	54.7	± 0.4
HLQ_λ	4	(0,0,0.995,0.01,0.8)	43.4	± 0.5	50.6	± 0.5	53.6	± 0.5	***55.3***	***± 0.5***	***55.6***	***± 0.5***
HLQ_λ	5	(0,0,0.995,0.005,0.9)	*40.7*	± *0.5*	*47.1*	± *0.5*	52.9	± 0.5	54.8	± 0.5	55.4	± 0.5
MC-AIXI	1	(50,8,3,0.85,0.3)	38.9	± 1.3	46.7	± 1.2	50.9	± 1.4	52.4	± 1.4	52.5	± 1.4
MC-AIXI	2	(50,16,2,0.95,0.9)	*39.6*	± *1.1*	*47.4*	± *1.2*	50.8	± 1.3	52.4	± 1.3	52.7	± 1.4
MC-AIXI	3	(100,8,3,0.9,0.6)	36.0	± 1.2	46.4	± 1.2	*51.6*	± *1.4*	*52.9*	± *1.4*	*54.3*	± *1.4*
MC-AIXI	4	(100,16,4,0.85,0.9)	***33.8***	***± 1.2***	42.6	± 1.1	47.2	± 1.3	49.3	± 1.3	*49.3*	± *1.4*
MC-AIXI	5	(100,32,3,0.05,1.0)	35.0	± 1.1	***40.2***	***± 1.1***	***44.0***	***± 1.2***	*46.6*	± *1.2*	49.7	± 1.2

Table 13: Estimated AIQ scores with a 0.95 confidence interval for all tested agent configurations. Agent's extremes in italics, overall extremes in bold. Agent's parameters explained in Section 2.6 (BF 5,3 Reference Machine).

Agent	**Configuration**		**AIQ Scores with Confidence Intervals for Episode Length**									
	#	*Parameters*	*1,000*		*3,000*		*10,000*		*30,000*		*100,000*	
freq	1	(0.03)	38.8	±0.5	*41.2*	*±0.5*	*43.5*	*±0.5*	*45.8*	*±0.5*	48.5	±0.5
freq	2	(0.05)	38.7	±0.5	41.8	±0.5	44.7	±0.5	46.7	±0.5	*49.0*	*±0.5*
freq	3	(0.07)	*39.1*	*±0.5*	41.8	±0.5	44.6	±0.5	*47.2*	*±0.5*	48.9	±0.5
freq	4	(0.09)	39.0	±0.4	*42.0*	*±0.5*	*44.9*	*±0.5*	46.7	±0.5	48.4	±0.5
freq	5	(0.11)	*38.6*	*±0.4*	41.9	±0.4	44.4	±0.4	46.4	±0.5	*47.9*	*±0.5*
Q_0	1	(0,0,0.5,0.04,0.6)	42.0	±0.4	44.7	±0.4	*45.4*	*±0.4*	*45.8*	*±0.4*	*46.0*	*±0.4*
Q_0	2	(0,0,0.5,0.03,0.7)	*42.4*	*±0.4*	46.1	±0.4	47.6	±0.4	48.1	±0.4	48.2	±0.4
Q_0	3	(0,0,0.5,0.02,0.8)	41.9	±0.4	*46.6*	*±0.4*	49.5	±0.4	50.4	±0.4	50.5	±0.4
Q_0	4	(0,0,0.5,0.01,0.9)	39.5	±0.5	44.3	±0.5	*49.9*	*±0.5*	*52.4*	*±0.4*	53.2	±0.4
Q_0	5	(0,0,0.5,0.005,0.95)	*38.6*	*±0.5*	*41.0*	*±0.5*	47.7	±0.5	52.2	±0.5	*54.0*	*±0.5*
Q_λ	1	(0,0.5,0.5,0.04,0.6)	44.4	±0.4	46.6	±0.4	*47.5*	*±0.4*	*47.8*	*±0.4*	*47.9*	*±0.4*
Q_λ	2	(0,0.5,0.5,0.03,0.6)	*44.6*	*±0.4*	47.3	±0.4	48.6	±0.4	48.9	±0.4	49.0	±0.4
Q_λ	3	(0,0.5,0.5,0.02,0.8)	44.0	±0.5	*48.2*	*±0.4*	51.0	±0.4	51.9	±0.4	51.9	±0.4
Q_λ	4	(0,0.5,0.5,0.01,0.9)	41.7	±0.5	46.3	±0.5	*51.3*	*±0.5*	*53.3*	*±0.5*	54.0	±0.5
Q_λ	5	(0,0.5,0.5,0.005,0.95)	*41.1*	*±0.5*	*43.3*	*±0.5*	48.9	±0.5	52.9	±0.5	*54.7*	*±0.5*
HLQ_λ	1	(0,0,0.99,0.02,0.7)	48.6	±0.4	50.9	±0.4	*51.9*	*±0.4*	*52.0*	*±0.4*	*52.0*	*±0.4*
HLQ_λ	2	(0,0,0.95,0.04,0.7)	***48.7***	***±0.4***	51.9	±0.5	53.2	±0.5	53.8	±0.5	53.7	±0.5
HLQ_λ	3	(0,0,0.99,0.04,0.6)	47.7	±0.5	***52.1***	***±0.5***	54.5	±0.5	55.1	±0.5	55.2	±0.5
HLQ_λ	4	(0,0,0.995,0.01,0.8)	44.1	±0.5	50.6	±0.5	***54.6***	***±0.5***	***55.9***	***±0.5***	56.4	±0.5
HLQ_λ	5	(0,0,0.995,0.005,0.9)	*41.8*	*±0.5*	*47.9*	*±0.5*	53.3	±0.5	55.6	±0.5	***56.5***	***±0.5***
MC-AIXI	1	(50,8,3,0.85,0.3)	***1.7***	***±0.5***	***2.3***	***±0.7***	***5.0***	***±0.9***	***5.8***	***±1.0***	***7.7***	***±1.1***
MC-AIXI	2	(50,16,2,0.95,0.9)	*37.5*	*±1.0*	*45.7*	*±1.3*	50.0	±1.4	51.8	±1.5	51.9	±1.5
MC-AIXI	3	(100,8,3,0.9,0.6)	3.4	±0.5	7.5	±0.6	9.6	±0.8	12.9	±0.9	13.0	±1.0
MC-AIXI	4	(100,16,4,0.85,0.9)	30.5	±0.9	42.6	±1.2	*50.2*	*±1.4*	*53.0*	*±1.4*	*53.9*	*±1.5*
MC-AIXI	5	(100,32,3,0.05,1.0)	33.6	±0.9	42.6	±1.1	47.1	±1.3	48.3	±1.3	49.5	±1.3

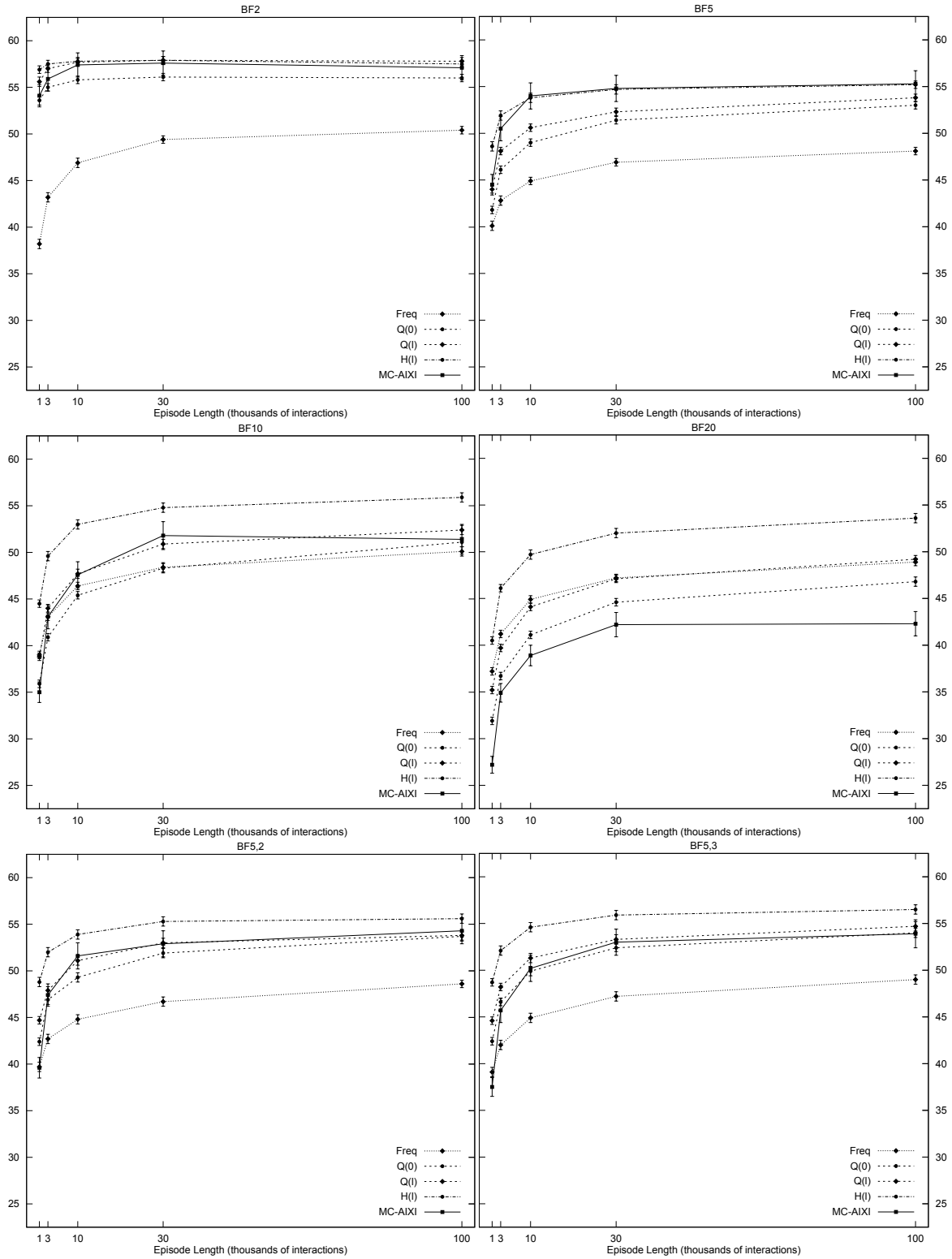

Figure 9: Best achieved estimated AIQ scores of agents as a function of episode length on BF 2, BF 5 (for comparison), BF 10, BF 20, BF 5,2, and BF 5,3 reference machines.

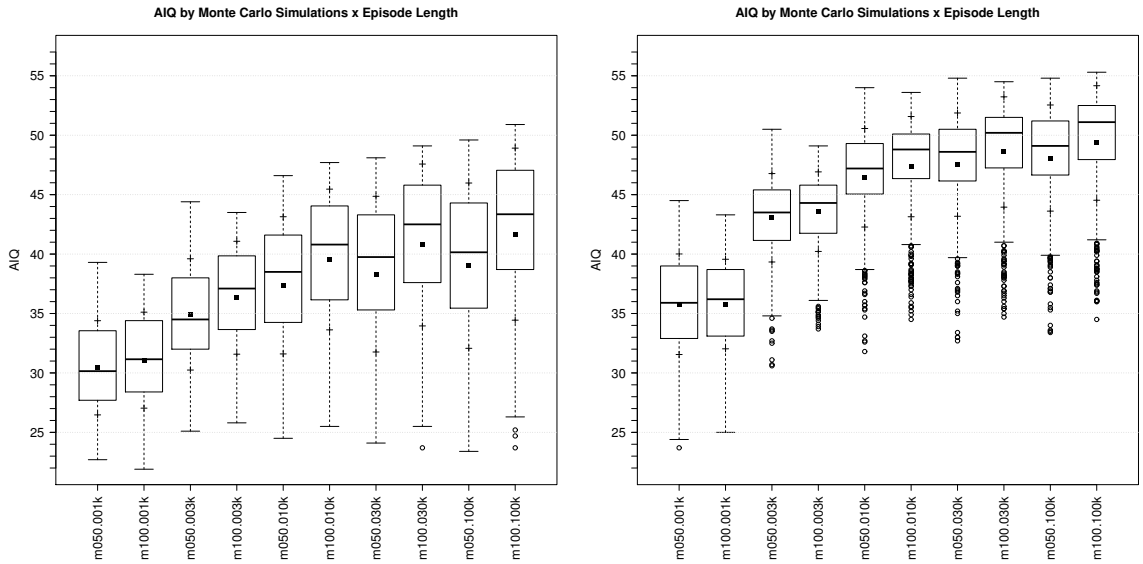

Figure 10: The impact of varying the *Monte Carlo simulations* parameter on AIQ of *MC-AIXI* configurations without exploration decay (to the left), and configurations with exponentially decayed exploration (to the right).

Appendix B. Explorative Analysis of MC-AIXI Results

Since in the case of *MC-AIXI* a limited parameter sweep was conducted, a more detailed results analysis can be attempted than in the case of other agents. As can be seen from Table 8 and Figure 2, there is a rather pronounced difference between the maximal and minimal AIQ achieved for a given episode length (especially noticeable with configurations not featuring exploration decay). Distribution of achieved AIQ is somewhat eccentric, favoring the higher values. There is a striking difference between the configurations with exploration decay and those without. Using exponential decay of exploration leads to generally better performance, both in terms of higher mean, median, quartiles, minimum, and maximum values, as well as less spread out results. Let us now examine the degree to which the parameters of *MC-AIXI* influence its results:

- Figure 10 shows the impact of using 50 versus 100 *Monte Carlo simulations* in the experiments, which is rather slight. Surprisingly, while mean and median scores are better for 100 simulations, maximal scores are better for 50 simulations at shorter episode lengths.

- Figure 11 shows the influence of setting *context tree depth* to 8, 16, and 32, which is also rather limited. If exploration decay is used, mean and median scores are somewhat better for the depth of 8 than for others, and overall are also notably less spread, except for at the shortest episode length.

- Figure 12 illustrates the impact of adjusting the *search horizon*, which is rather pronounced. Setting the search horizon to 1 leads to rather low AIQ and rather spread out results across all episode lengths, while setting it to 2 or 3 increases the AIQ by a notable margin. A search

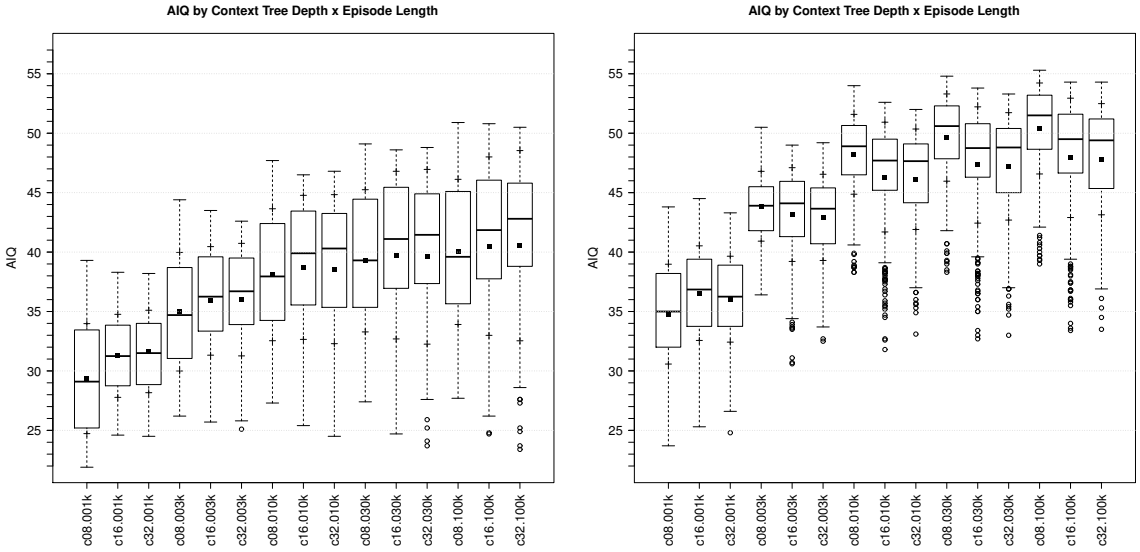

Figure 11: The impact of varying the *context tree depth* parameter on AIQ of *MC-AIXI* configurations without exploration decay (to the left), and configurations with exponentially decayed exploration (to the right).

horizon value of 4 and 5 gives rather low AIQ at short episode lengths, however, when the number of iterations is increased it improves notably.

- Figure 13 illustrates the influence of varying the *exploration* parameter, which seems to further depend on whether it is decayed or not. In the case of exponentially decayed exploration, the influence is minimal, while in the case of undecayed exploration, it is rather pronounced. Increasing the exploration in the second case decreases the resulting AIQ for all episode lengths.

- Figure 14 shows the impact of varying the *exploration decay*, which is noticeable at lower episode lengths, where configurations that decayed more slowly (with decay of 0.99, and 0.995) perform rather more poorly. However, the effect declines with increasing episode length until the means and medians differ only slightly. Also, the results of configurations with faster decay rates get more spread out at higher episode lengths.

Consequently, there seems to be some influence of the chosen parameter values on the *MC-AIXI* performance in the AIQ test. However, the nature of the influence is not a simple one, and therefore further analysis is needed. To model the impact of *MC-AIXI* parameters on its AIQ, data mining methods were used, specifically the classification and regression trees (CART) of Breiman et al. (1984) and the conditional inference trees (CIT) by Hothorn, Hornik, and Zeileis (2006).

For the analysis, AIQ was used as a dependent variable for regression, while for classification, a derived dependent variable AIQ_{rcat} was introduced with ordered values low (AIQ below 1st quartile for given EL), medium (AIQ between 1st and 3rd quartiles for given EL), and high (AIQ above 3rd quartile for given EL). Since AIQ is a mean over multiple tested environment programs, looking

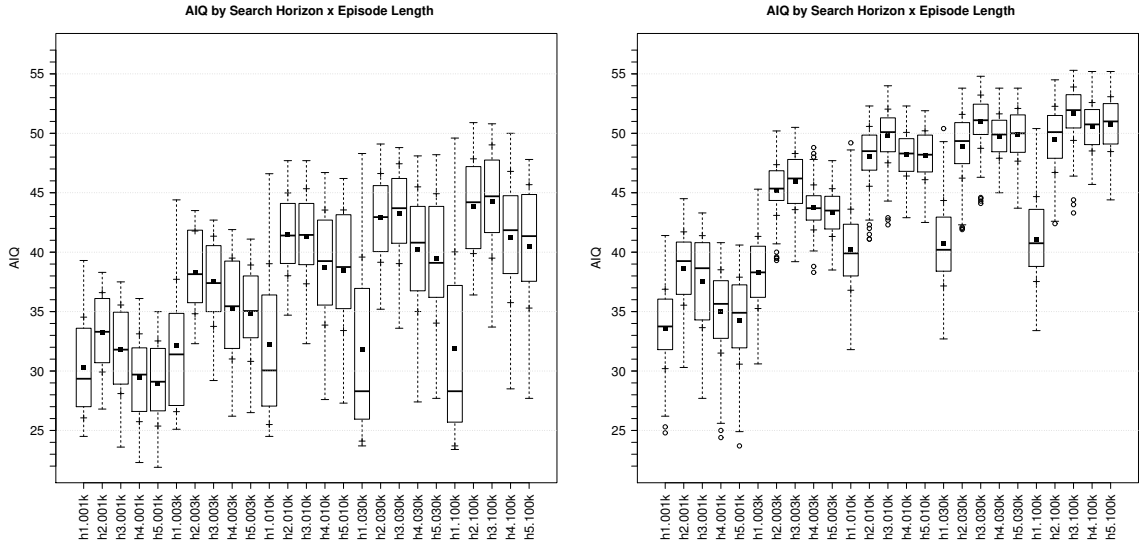

Figure 12: The impact of varying the *search horizon* parameter on AIQ of *MC-AIXI* configurations without exploration decay (to the left), and configurations with exponentially decayed exploration (to the right).

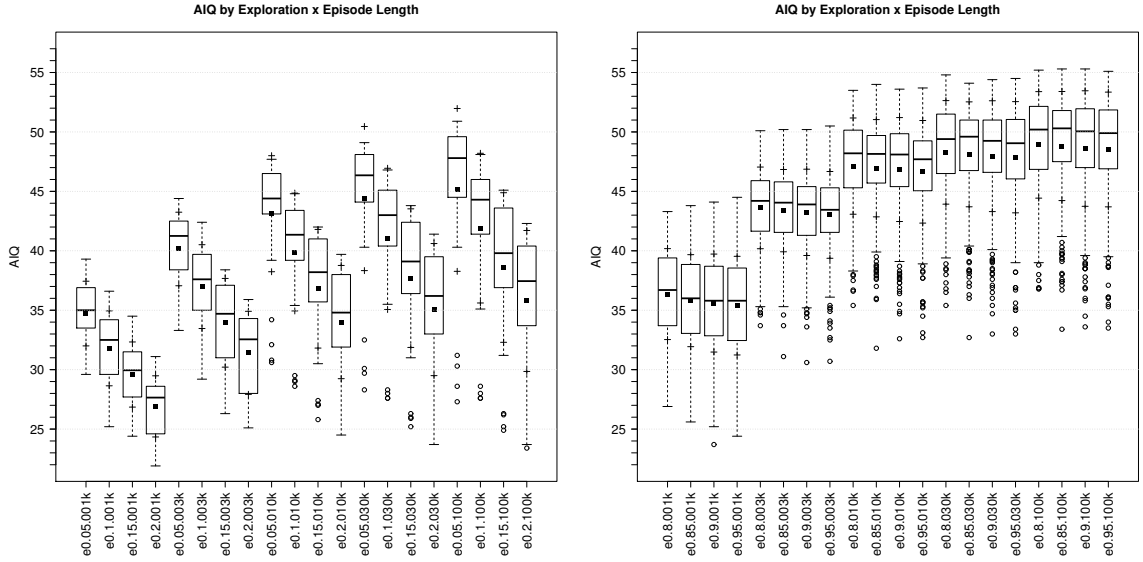

Figure 13: The impact of varying the *exploration* parameter on AIQ of *MC-AIXI* configurations without exploration decay (to the left), and configurations with exponentially decayed exploration (to the right).

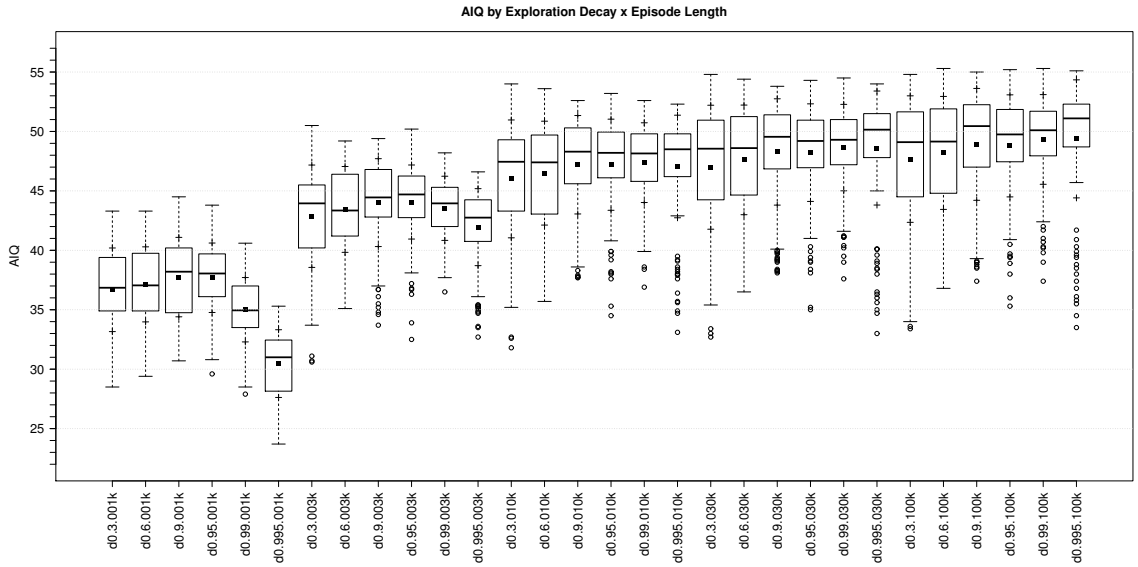

Figure 14: The impact of varying the *exploration decay* parameter on *MC-AIXI* AIQ results (only configurations with enabled exponential decay shown).

closer at its standard deviation (SD) might provide further insight into the possible differences in the spread of performance of *MC-AIXI* configurations among tested environment programs. This is especially interesting when considered together with the AIQ score so that configurations with relatively high AIQ, but low SD could be identified. SD was first classified as low, medium, or high using rules analogous to the case of AIQ_{rcat}, then a combined dependent variable $\text{AIQSD}_{\text{rcat}}$ was introduced with ordered values: low;high (lh), low;medium (lm), medium;high (mh), low;low (ll), medium;medium (mm), high;high (hh), medium;low (ml), high;medium (hm), and high;low (hl) where the first part is AIQ classification and the second one is SD classification. This classification, however, proved too fine-grained, with data fitting particularly the first and last category being very rare. Therefore, another derived dependent variable was introduced, $\text{AIQSD}_{\text{arcat}}$, with ordered values low (aggregating lh, lm, and mh), medium (aggregating ll, mm, and hh), and high (aggregating ml, hm, and hl). As covariates, all the manipulated *MC-AIXI* parameters were used: *number of Monte Carlo simulations* (MC), *context tree depth* (CTD), *search horizon* (AH), *exploration* (E), and *exploration decay* (ED), all with numerical values, as well as a derived Boolean attribute, *decay* (D), signifying whether exponential decay of exploration was used or not. The decisive episode length is 100,000 interactions since the score should be well converged by then. Additionally, in order to see whether some parameters are only influential at lower episode lengths, another analysis was attempted using results on all episode lengths with another covariate EL.

Regression trees were attempted both for EL of 100,000 interactions, as well as for results of all tested episode lengths. However, since there is a sizable impact of EL on the AIQ, higher EL was used as the main predictor for higher AIQ. Among other predictors for high AIQ, there was $\text{AH} \geq 2$, and $\text{D} = \text{true}$. At an EL of 1,000 interactions, configurations with $\text{ED} = 0.995$ were grouped with undecayed configurations as low performing as opposed to other values of ED. Configurations with $\text{AH} = 1$, $\text{D} = \text{true}$, and $\text{CTD} > 8$ were identified as especially low performing at $\text{EL} > 1000$. In

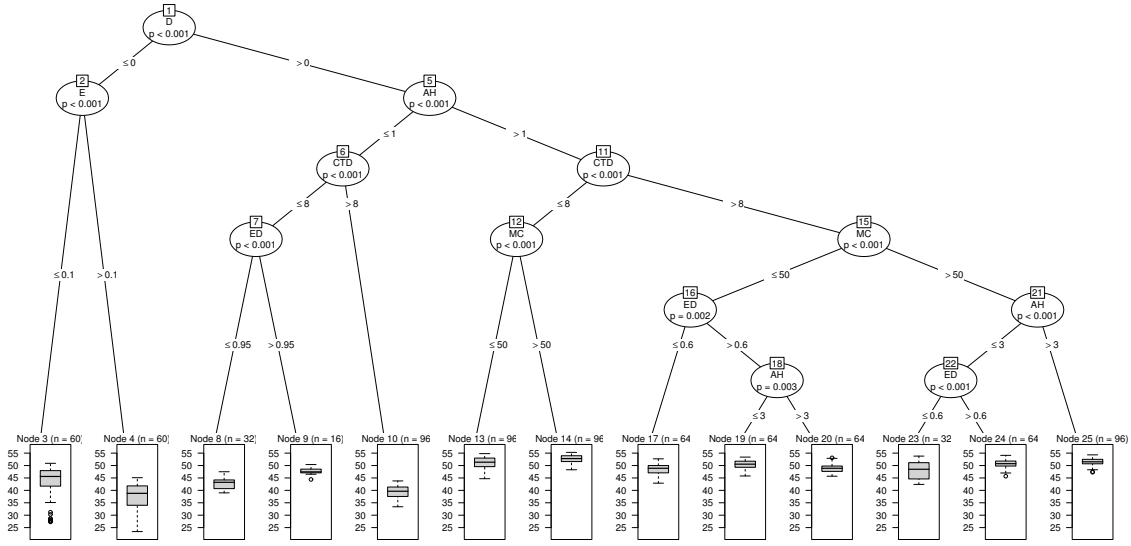

Figure 15: Conditional inference regression tree showing the impact of varying the parameters of *MC-AIXI* on its AIQ at EL of 100,000 interactions.

the analysis at EL of 100,000 interactions, AH ≥ 2 was confirmed as the main prognostic factor for high AIQ, followed by D = true and CTD $= 8$. Configurations with AH $= 1$, D = false, and CTD > 8 were identified as especially low performing.

Classification trees were also attempted both for EL of 100,000 interactions, as well as for all tested episode lengths to classify both AIQ_{rcat}, and $\text{AIQSD}_{\text{rcat}}$. However, the resulting models achieved a relatively low total percentage of correctly classified configurations.

Finally, conditional inference tree analysis was conducted. Since this method allows for a more detailed presentation of the resulting groups, and takes into account statistical properties of the data when constructing the tree (Hothorn, Hornik, and Zeileis, 2006), the results of this method were chosen to be presented in this paper in more detail:

- For a conditional inference regression tree showing the impact of *MC-AIXI* parameters on its AIQ at EL of 100,000 interactions see Figure 15. The tree identified *decay* as the main predicting attribute of AIQ. Configurations with D = false were further divided according to the parameter *exploration*: the group with E ≤ 0.1 was significantly better than the group with higher *exploration*. The main effort of the tree, however, focused on the configurations with D = true. There, the main contributing factor was found to be *search horizon*, followed by *context tree depth*. Configurations with D = true, AH $= 1$, CTD $= 8$, and MC $= 100$ were identified as especially high performing (Node N#14), followed by configurations with D = true, AH ≥ 4, CTD ≥ 16, and MC $= 100$ (Node N#25), and those with D = true, CTD ≥ 16, MC $= 100$ and AH of 2 or 3 (Node N#24). Meanwhile, the configurations with D = true, and AH $= 1$, were identified as rather poorly performing (Nodes N#8, and N#10) with a notable exception of those with CTD ≥ 16, and ED ≥ 0.99 (Node N#9).

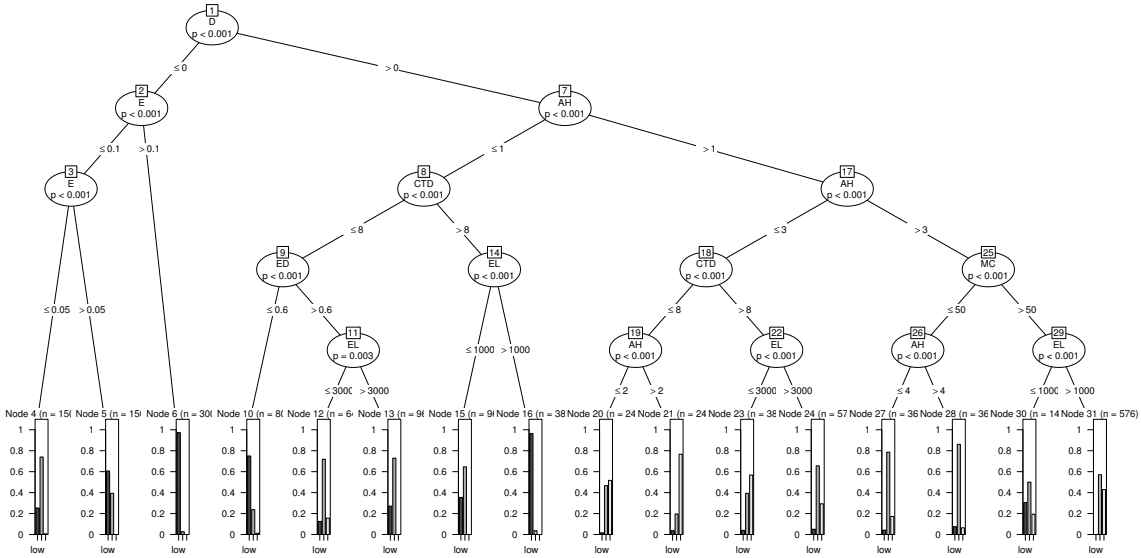

Figure 16: Conditional inference classification tree showing the impact of varying the parameters of *MC-AIXI* on its AIQ at all test episode lengths.

- A conditional inference classification tree showing the influence of *MC-AIXI* parameters on its AIQ_{rcat} at all test episode lengths is depicted in Figure 16. The tree also identified *decay* as the main predicting factor of AIQ_{rcat}. Configurations with D = false were further partitioned into three groups according to their *exploration*: $\text{E} > 0.1$ with mostly low AIQ_{rcat} (Node N#6), $\text{E} = 0.05$ with mid-range AIQ_{rcat} (Node N#4), and $\text{E} = 0.1$ somewhat between medium and low AIQ_{rcat} (Node N#5). For configurations with D = true, the *search horizon* attribute was also identified as the main predictor, followed by *context tree depth*. Configurations with D = true, AH = 3, and CTD = 8 (Node N#21) were identified as mostly having high AIQ_{rcat}, followed by configurations with D = true, CTD = 8, AH of 2 or 3 at $\text{EL} \leq 3,000$ (Node N#23), and those with D = true, AH = 2, and CTD = 8 (Node N#20) which were, in both cases, somewhere between medium and high AIQ_{rcat}. Configurations with D = true, and AH = 1 were classified as mostly medium (Nodes N#12, N#13, and N#15), except from those with $\text{CTD} \geq 16$ at $\text{EL} \geq 3,000$ (Node N#16), and those with CTD = 8, and $\text{ED} \leq 0.6$ (Node N#10) which were classified as having mostly low AIQ_{rcat}.
- Conditional inference classification trees for the impact of *MC-AIXI* parameters on its $\text{AIQSD}_{\text{rcat}}$, and $\text{AIQSD}_{\text{arcat}}$ for all the tested episode lengths ended up being rather complex. Attempts at simplifying them resulted in a rather balanced distribution of values in terminal nodes. For the sake of brevity, only the selected terminal nodes of the large trees are shown in Figure 17. The following configurations were classified as rather highly performing with relatively high AIQ but relatively low SD:
  - D = true, MC = 100, CTD = 32, AH > 3, E = 0.8, and ED = 0.3 (a);
  - D = true, MC = 100, CTD = 8, AH > 1 ≤ 3, and ED ≤ 0.6 (b);
  - D = true, MC = 50, CTD = 8, AH = 3, and ED ≤ 0.95 (c);

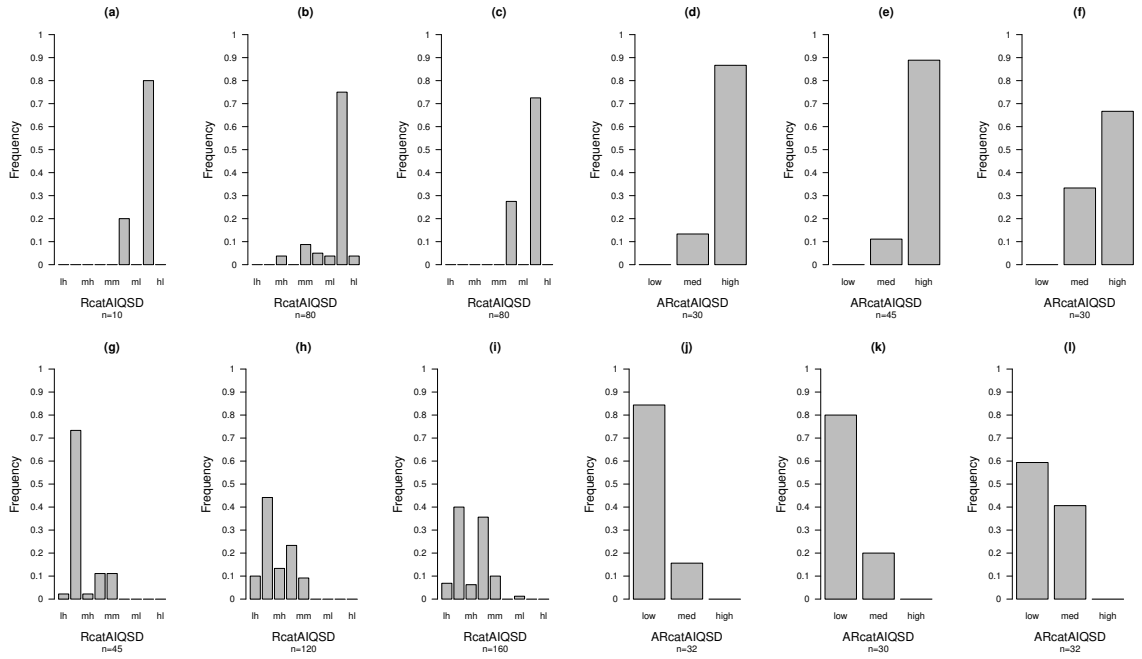

Figure 17: Chosen terminal nodes of a conditional inference classification tree showing the impact of varying the parameters of *MC-AIXI* on its AIQ and SD at all tested episode lengths.

- D = true, MC = 100, CTD $\leq$ 16, AH = 2, E $\geq$ 0.85, and ED = 0.3 (d);
- D = true, MC = 50, AH = 3, E $\geq$ 0.85, and ED = 0.9 (e);
- D = false, MC = 50, AH = 3, and E $\leq$ 0.1 (f);

Meanwhile, the following configurations were identified as rather poorly performing with relatively low AIQ but relatively high SD:

- D = true, MC = 50, AH = 1, E $\geq$ 0.85, and ED = 0.9 (g);
- D = true, MC = 100, AH = 1, and ED $\leq$ 0.6 (h);
- D = true, MC = 100, CTD $\geq$ 16, AH = 1, and ED $\geq$ 0.9 (i);
- D = true, CTD = 16, AH = 1, and ED = 0.995 at EL $\geq$ 3,000 (j);
- D = true, MC = 100, CTD $\leq$ 16, AH = 1, E = 0.8, and ED $\leq$ 0.9 (k);
- D = true, CTD = 32, AH = 1, and ED = 0.995 at EL $\geq$ 3,000 (l);

Appendix C. Extended AIQ Test

As a part of the work on this paper, the AIQ test was extended by adding the following functionality:

- The parameter *exploration decay* of *MC-AIXI* agent can be set.
- Standard deviation (SD) of AIQ score is printed by `ComputeFromLog.py`.
- A sample of all syntactically valid programs can be generated by `BF_Sampler.py` using `--theoretical_sampler` option.
- Environment programs actually used by the AIQ test can be saved by `AIQ.py` using `--save_samples` option.
- The current AIQ estimate can be saved every 1,000 interactions by `AIQ.py` using the option `--verbose_log_el`.
- A sample of environment programs with a given minimal length can be generated by `BF_Sampler.py` using `-l minimal_length` option. By default, programs shorter than the specified minimal length are dropped. To use the method that extends shorter programs, add option `--extend_shorter`.

Full sources are available from: `https://github.com/xvado00/AIQ/archive/v1.1.zip`.