An Information-Theoretic Predictive Model for the Accuracy of AI Agents Adapted From Psychometrics

Nader Chmait, David L. Dowe, Yuan-Fang Li, and David G. Green

Faculty of Information Technology, Monash University, Clayton, Australia

Abstract. We propose a new model to quantitatively estimate the accuracy of artificial agents over cognitive tasks of approximable complexities. The model is derived by introducing notions from algorithmic information theory into a well-known (psychometric) measurement paradigm, Item Response Theory (IRT). A lower bound on accuracy can be guaranteed with respect to task complexity and the breadth of its solution space using our model. This in turn permits formulating the relationship between agent selection cost, task difficulty and accuracy as optimisation problems. Further results indicate some of the settings over which a group of cooperative agents can be more or less accurate than individual agents or other groups.

1 Introduction and Background

Turing's imitation game [31] inspired a range of attempts to measure the intelligence of artificial agents. More recently, a formal (machine) intelligence test [9] consisting of sequence-completion exercises was devised. Later, fuzzy integrals were used [1] to measure intelligence in machines by calculating a Machine Intelligence Quotient. Shortly after, a simple computer program that succeeded in passing a variety of IQ tests was presented [25], raising questions on the appropriateness of intelligence tests for machine assessment. After the definition of *universal intelligence* [17], many (algorithmic) information-theoretic studies were put forward to formally quantify the intelligence of individual AI agents [11, 12] as well as AI collectives [4].

Independently, a series of measurement theories have been proposed in psychometrics and applied to human intelligence. One of the earliest milestones in human intelligence testing was Thurstone's letter series completion problems [30] and, more recently, Raven's Progressive Matrices test [23] which recorded strong correlation with Spearman's general intelligence factor [29]. More general tests consisting of a variety of evaluation tasks were developed, and they came to be known as "Intelligence Quotient" or simply IQ tests. Examples of such tests are the Stanford-Binet test [24] and the Wechsler intelligence scales for adults and children [32]. Another mainstream achievement in psychometrics was the development of Item Response Theory (IRT) [21], also referred to as latent trait theory. IRT is among the most popular measurement classes used in psychometrics for evaluating traits, or abilities, and producing accurate rankings from test scores, by applying mathematical models to testing data. In the context of IRT, a trait or an ability might be physical or psychological (cognitive and non-cognitive, e.g., a personality or behavioural characteristic) [5]. Recently, IRT was successfully adopted to analyse machine learning models by providing an instance-wise analysis of a series of datasets and classifiers [22]. In this paper, we show how to adapt models from psychometrics and IQ tests, based on notions from algorithmic information-theory, to artificial intelligence in order to estimate the (cognitive) abilities of artificial agents and predict their accuracies.

2 Motivation and Main Contributions

Advances in psychometrics are not yet thoroughly applied for predicting the accuracy of AI agents despite their success in evaluating human cognitive abilities. While the AI discipline adheres to the mainstream concept of intelligence [8], general IQ tests might not be appropriate in their current form for evaluating machine intelligence [6]. In fact, even test batteries that might be suitable for practically evaluating AI (and knowledge based systems [14]) show some caveats. For instance, such tests measure an average performance (of one or more abilities) of AI agents over a set of tasks or environments but it is ambiguous how the results from these tests can be used to predict the accuracy of an agent over a particular task complexity without actually administering that task to the agent. In addition to many theoretical studies discussed in [10], empirical studies such as [4, 3] demonstrated that task complexity and the breadth of its solution space are major factors influencing the performance of artificial agents. Hence, quantitatively predicting the accuracy of artificial agents across different task complexities and solution spaces is clearly an important feature that has not been addressed so far. Furthermore, intelligence test scores can be unreliable since agents usually exhibit non-uniformity between their performances over different problems/settings. This has implications for selecting agents to solve tasks, particularly when there is cost (e.g., processing time) associated with utilising agents, and understanding the collective accuracy of cooperative agents of different (cognitive) abilities.

By merging notions from both psychometrics and (algorithmic) information theory, we develop a hybrid model to quantitatively estimate the accuracy of AI agents over tasks of measurable complexities. We demonstrate its functionality over a class of prediction and inference problems as this class is considered as reflecting some of the principal traits of intelligence both in psychometrics [8] and artificial intelligence [20, 7, 10]. Using the predictive model, we show how to identify agents that can guarantee a lower bound on accuracy with respect to task complexity and the breadth of its solution space. We analyse settings over which a group of (voting) agents can be more or less effective than individual agents, or other groups, and identify circumstances that can be counterintuitive to the conclusions drawn from intelligence tests. In the next section we outline important properties and constraints that our model needs to embrace.

3 Desirable Properties for Assessment

Given a subject (cognitive agent) to be evaluated over a task/problem:

1. The model must return a *quantitative* measure (on an interval scale) of the estimated subject's accuracy over this task without the need to administer it to the subject.
2. The accuracy of a subject (its probability of success in solving a task) predicted by the model is expected to be proportional to its (relevant cognitive) ability over that task, and inversely proportional to the difficulty of the task.
3. In order to conform to the *limiting* behaviour of real agents, the model should use the asymptotic minimum (p_{rand}: probability of correctly selecting a random solution from the sample space) as a lower-bound on accuracy.
4. The model should be applicable over different tasks of measurable difficulties.
5. The difficulty measure should be general enough to accommodate a wide range of tasks.
6. The model should be applicable to different agent types and cognitive systems.

Earlier information-theoretic studies on (artificial) intelligence [11, 13] and inductive-inference [28, 18] discussed (among others) two general dimensions of task difficulty, (i) Shannon's entropy [26] which is related to the uncertainty and breadth of the solution search space, and (ii) the algorithmic information-theoretic (in particular the Kolmogorov) complexity [15, 20] of the task. We take into account both dimensions of difficulty in the design of our model.

4 A Predictive Model of Agent Accuracy

Inspired by the 2-parameter logistic model [2] of IRT [21], we propose a mathematical model for predicting a subject's expected accuracy on a given task/problem of measurable complexity.

Definition 1. *Let x denote a task/problem of a (theoretical) difficulty $\mathcal{D}$ such that the solution to x belongs to the alphabet (or solution space) $S=\{s_1,s_2,...,s_m\}$. We define (an estimate of) the accuracy of an agent with ability $\alpha \in \mathbb{R}^+$ over that task to be:*

$$P_{\mathcal{D},\alpha,m} = \frac{1}{m} + e^{-\frac{\mathcal{D}}{\alpha}} \cdot \left(1 - \frac{1}{m}\right) \tag{1}$$

which corresponds to the probability of that agent guessing the correct solution to x.

The above model has the following important properties. For a given task of a (hypothetically) negligible difficulty, the probability of solving this task is $\lim_{\mathcal{D}\to 0} P_{\mathcal{D},\alpha,m} = 1$. The probability $P_{\mathcal{D},\alpha,m}$ of a subject with ability $\alpha > 0$ solving a task is (exponentially) proportional to the subject's ability, and inversely proportional to the difficulty of the task $\mathcal{D}$, and the breadth of its solution space $m \in \mathbb{N}^+$. Moreover, when task difficulty $\mathcal{D}$ is very high relative to α (or when the subject's ability α is small), the probability of success $P_{\mathcal{D},\alpha,m}$ converges to a random guess equivalent to $1/m$, which is the asymptotic minimum[1]. For instance, on a binary test problem (e.g., coin toss problem with S={Heads, Tails}) with $m = 2$, an agent with ability α has an accuracy $P_{\mathcal{D},\alpha,m} = 0.5 + e^{\frac{-\mathcal{D}}{\alpha}}(0.5)$. When the ability α is close to zero, $P_{\mathcal{D},\alpha,m} \approxeq 0.5$. For many problems, the theoretical task difficulty $\mathcal{D}$ can be derived from the simplest solution (policy) to the task, and therefore can sometimes be linked to the *complexity* of the (description of the) task, or the complexity of the description of its policy. Consequently, the difficulty of the task can be linked to its Kolmogorov complexity [15, 20]. Since the Kolmogorov complexity is uncomputable, methods like Levin's Kt complexity [20, 19] or the Lempel-Ziv (compression) algorithm [18] can be used as practical alternatives (to bound it and possibly approximate it). For the rest of this paper, we will use the Kolmogorov complexity of the task as a derivation of its (theoretical) difficulty. The suggested model returns the probability of a subject solving a given task of a measurable complexity as a function of its (previously measured) ability. The ability could be defined as a vector of weighted atomic sub-abilities s.t. α is a linear combination of $[w_1\alpha_1 + w_2\alpha_2 + ... + w_t\alpha_t]$. The model in Eq. 1 is a simple case of the latter where, for some integer $z \leq t$, the ability $\alpha = w_z\alpha_z$ and $\sum_{j=1,j\neq z}^{t} w_j = 0$ in $[w_1\alpha_1 + w_2\alpha_2 + ... + w_t\alpha_t]$.

We will use a formal intelligence test from the literature of AI, the C-test [9], to measure an agent's ability α over a class of tasks. $\mathcal{D}$ and m are input parameters to the model typically being measured by some earlier assessment or derived directly from the problem. We refer to the model defined in Eq. 1 as the *IRT model* for brevity, and use the terms *accuracy* and *performance* alternately (only) as measures of the probability of success at solving a (cognitive) task.

[1] For simplicity and without loss of generality, $1/m$ is used in Eq. 1 to replace the probability p_{rand} of an agent randomly guessing (one of) the correct solutions to the problem.

5 Assessing Inference Abilities

The C-test [9] is a compression-based intelligence test that measures the ability of a subject doing inductive-inference and finding the best explanation for sequences of various complexities. It reflects the *fluid* intelligence of the evaluated subject. The idea is to record the performance of a subject over a series of patterns of increasing incomprehensibilities (or complexities). The complexity of a C-test sequence is formally measured using Levin's Kt complexity [20] as a practical alternative to (and possibly a rough bound on) its Kolmogorov complexity. Given $\Sigma = \{a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z\}$, and a sequence θ of length m where each $\theta_i \in \Sigma$, the task consists of predicting the next letter $\theta_{m+1} \in \Sigma$ which correctly completes the sequence. Given a C-test consisting of a collection of test sequences $CT = (seq_1,...,seq_n)$ with their corresponding answers (solutions) $S = (\theta^1_{m+1},...,\theta^n_{m+1})$ and corresponding complexities $K = (k_1,...,k_n)$, the average score $\tilde{r}$ of an agent π with guesses $S' = (\theta'^1_{m+1},...,\theta'^n_{m+1})$ over CT is: $\tilde{r} = \frac{1}{\sum_{z=1}^{n} k_z} \cdot \sum_{z=1}^{n} k_z \times hit(\theta'^z_{m+1}, \theta^z_{m+1})$, where the function $hit(a,b) \leftarrow \begin{cases} 1 & \textit{if } a == b \\ 0 & \textit{otherwise} \end{cases}$, and the complexity of the sequence k_z is used as a weight in order to give more importance to more difficult questions. The C-test score will be used to determine the inductive-inference ability α of a subject, further used as a parameter in the model (Eq. 1). The reasons for selecting the C-test are, firstly, the test by definition measures an (inductive inference related) ability, in this case the ability of finding the best explanation for a given sequence using induction. The test is well formulated and is exclusively defined in computational terms. It generates sequences (tasks) within a range of complexities $7 \leq D \leq 15$, using Levin's Kt approximation [9] (as a practical alternative to *Kolmogorov* complexity). The C-test results are highly correlated with those from classical psychometric (IQ) tests [9]. The test sequences are formatted and presented in a quite similar way to psychometric tests. Hence, the test can be applied to machines in the same way it is applied to humans. There is typically one exclusive correct (simplest) answer for any of the test sequences, making the results uncoincidental and representative of the testee's accuracy.

Measuring abilities: Table 1 holds the definitions of a few agent behaviours to be evaluated over the C-test. Their scores are used to measure their (inductive inference) ability α and are plotted in Fig. 1 along with their corresponding accuracies $P_{\mathcal{D},\alpha,m}$ generated using the IRT model (Eq. 1). More advanced algorithms for sequence prediction problems exist but since the choice of agent behaviours is not particularly relevant to the validity of the model we restrict our selection to those in the Table 1. The agents' abilities were calculated as a function of their C-test scores using $\alpha = \omega\tilde{r}$, where α is the ability of agent π with score $\tilde{r}$, and $\omega \in \mathbb{R}$ is a fitting parameter selected in such a way to (i) ensure that the agent's moderate accuracy, of $0.5(\max P_{\mathcal{D},\alpha,m} + \min P_{\mathcal{D},\alpha,m}) \equiv 0.5(1 + 1/m)$, falls under the area of *discriminative* task complexities $\int_{\mathcal{D}=6}^{\mathcal{D}=16} P_{\mathcal{D},\alpha,m}$ (following [9]) and, (ii) minimise the mean squared error between the IRT model and C-test scores. Our model nicely illustrates the agents' average accuracies as illustrated in Fig. 1 despite the large non-uniformity in their behaviours and performances.

6 Predicting Agent Performance

While results from the C-tests are all alone interesting, we have no means to extrapolate them or predict the agent performances over different sequence complexities and solution space sizes without re-running the test. However, the expected accuracies of an agent can easily be

Table 1: Sample agent behaviours evaluated over the C-test.

Random agent: given a sequence *seq*, a *random agent* π^{rand} randomly uniformly selects a letter from Σ and returns it as its answer θ'_{m+1} (Refer to Sec. 5).

Mode agent: given a sequence *seq*, a *mode agent* π^{mode} looks for the most repeated or frequent letter(s) in *seq* to predict the next letter. If more than one letter satisfy the criteria, it chooses the left-most one appearing in the sequence.

Min-repetition agent: given a sequence *seq*, a *min-repetition agent* π^{mr} looks for the least repeated letter in *seq* to predict the next letter.

Min-distance agent: given $seq = (\theta_1, \theta_2, \ldots, \theta_m)$, agent π^{mind} looks for the minimal alphabetical distance (Def. 2) between all consecutive letters of *seq* and infers the next letter θ'_{m+1} by adding this distance to *seq*'s last letter θ_m.

Definition 2. *The alphabetical distance $d(\gamma - \beta)$ between two characters β and γ in an alphabet Σ is equal to the difference between their index positions in the totally ordered set $(\Sigma, \leq)$ in* $\bmod |\Sigma|$.

For instance, the distance between any two consecutive letters in the alphabet is 1, and the distance between the first character a and the last one z is equal to $d(z-a) = 26 - 1 = 25$. So, given a C-test sequence $seq = (\theta_1, \theta_2, \ldots, \theta_m)$, agent π^{mind} calculates the distance $d^i := d(\theta_{i+1} - \theta_i)$ following Definition 2 between two consecutive elements of *seq* for all $i \in \{1, \ldots, m-1\}$ returning a pattern (list) of distances $D = (d^1, d^2, \ldots, d^{m-1})$. Then, π^{mind} looks for the minimal alphabetical distance $d^{min} \in D$ as follows: $d^{min} \leftarrow \text{argmin}_{d \in D}\ freq(d, D)$ where $freq(d, D)$ is a function that returns the rate at which d occurs in $\mathcal{D}$. Agent π^{mind} finally chooses $\theta'_{m+1} \in \Sigma$ such that $d(\theta'_{m+1} - \theta_m) = d^{min}$.

Max-distance agent: this is the opposite behaviour of *min-distance agent*. Given a sequence $seq = (\theta_1, \theta_2, \ldots, \theta_m)$, a *max-distance agent* π^{maxd} calculates the distance $d^i := d(\theta_{i+1} - \theta_i)$ between the consecutive elements of *seq* for all $i \in \{1, \ldots, m-1\}$ returning a pattern (list) of distances $D = (d^1, d^2, \ldots, d^{m-1})$. It then looks for the maximal alphabetical distance: $d^{max} \in D \leftarrow \text{arg max}_{d \in D}\ freq(d, D)$ (from above definition). It finally chooses $\theta'_{m+1} \in \Sigma$ such that $d(\theta'_{m+1} - \theta_m) = d^{max}$.

Pattern agents: a pattern agent π^{pt} looks for a repeating distance pattern between the elements of *seq* and completes it to infer θ'_{m+1}. To implement this behaviour, the problem is divided into $m-1$ tasks $\{t_1, t_2, \ldots, t_{m-1}\}$ assigned to agents $\{\pi_1^{pt}, \pi_2^{pt}, \ldots, \pi_{m-1}^{pt}\}$ respectively. Agent π_y^{pt} calculates $d(\theta_{i+y} - \theta_i) \forall i \in \{1, \ldots, m-y\}$ and generates a list of distances $D_y = (d_y^1, \ldots, d_y^k)$ where $k = m - y$, and $d_y^i := d(\theta_{i+y} - \theta_i)$. Then, π_y^{pt} searches for the occurrences of the longest possible pattern in D_y and continues D_y by adding d_y^{k+1} following the pseudo-algorithm below.

Input: set of distances $D_y = (d_y^1, d_y^2, \ldots, d_y^k)$.

Output: next distance d_y^{k+1} in D_y.

Extract the unique elements of D_y.
Store elements in a list U_y in order of appearance.
Find the starting index for each substring occurrence U_y in D_y.
Store index in vector v.
if $|v| > 1$ **then**
    $P \leftarrow D_y(v(1) : v(2) - 1)$ ▷ $v(i)$ is the i'th element of v
else if $|v| \leq 1$ & $|D_y| > 1$ **then**
    $P \leftarrow D_y(|D_y| - 1)$
else
    $P \leftarrow D_y$
end if
$ind \leftarrow |D_y| - |P| \times |v|$
if $ind > 0$ **then**
    $d_y^{k+1} \leftarrow P(ind+1)$
else
    $d_y^{k+1} \leftarrow P(1)$
end if
return d_y^{k+1}

Finally, agent π_y^{pt} makes its guess θ'_{m+1} for the next letter of *seq* such that: $d(\theta'_{m+1} - \theta_{m+1-y}) = d_y^{k+1}$.

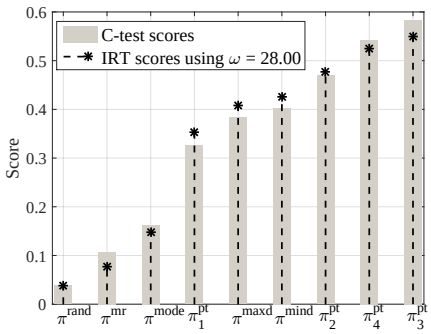

Fig. 1: Final C-test score $\tilde{r}$ of 9 different agents behaviours (defined in the Table 1) and their corresponding IRT accuracies taken from Eq. 1, using an $\alpha = \omega\tilde{r}$ s.t. $\omega = 28$.

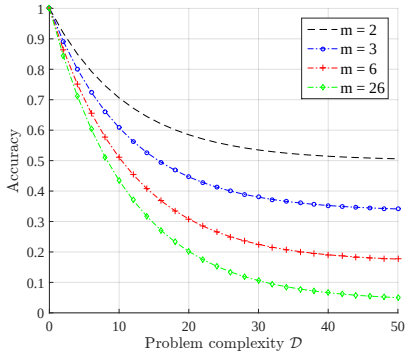

Fig. 2: IRT accuracy of agent π^{mind} with ability $\alpha = 11.28$ over inference tasks of different hypothetical (Kolmogorov) complexities D and problem solution space sizes m.

generated from the IRT model over inference tasks of different complexities. An example is illustrated in Fig. 2 showing the predicted accuracies of agent π^{mind} (refer to Table 1) across different hypothetical (Kolmogorov) complexities D and problem solution space sizes m. For any fixed difficulty $\mathcal{D}$, the IRT model shows that the difference in accuracy measures $P_{\mathcal{D},\alpha,m_1} - P_{\mathcal{D},\alpha,m_2}$ over two solution space sizes $m_2 > m_1$ is: $\frac{1}{m_1} + \frac{e^{\frac{-\mathcal{D}}{\alpha}}}{m_1} - \frac{1}{m_2} - \frac{e^{\frac{-\mathcal{D}}{\alpha}}}{m_2} = \frac{(1+e^{\frac{-\mathcal{D}}{\alpha}})(m_2 - m_1)}{m_1 \cdot m_2}$, meaning that this difference is greater over smaller $m \in \mathbb{N}^+$. This can also

be observed in Fig. 2. For consecutive values of m, $P_{\mathcal{D},\alpha,m} - P_{\mathcal{D},\alpha,m+1} = (1+e^{\frac{-\mathcal{D}}{\alpha}})/(m^2+m)$, and therefore, for very large m, any further increase in m has a negligible effect on the accuracy.

6.1 Relationship Between Accuracy and Difficulty

Figure 3 shows the shift in accuracies of a pool of example classifiers of hypothetical (classification) abilities $\alpha \in [1,8]$ across several $\mathcal{D}$ and m values. We observe that m has a greater influence than $\mathcal{D}$ on the accuracy of those classifiers with poor abilities $\alpha < 3$ and thus their scores are asymptotically bounded by $1/m$, while the opposite is true for more adept classifiers with stronger abilities. This type of analysis can be used to identify the minimal ability value for a classifier to be considered effective compared to, for example, a simple random classifier. One can further put a bound on the task complexity that an agent can solve with

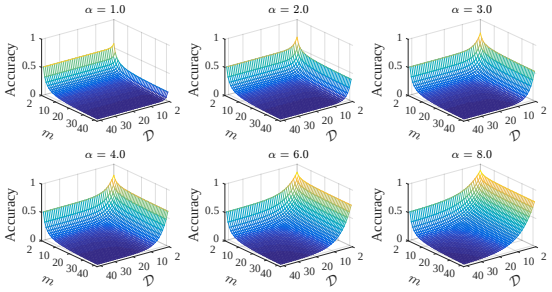

Fig. 3: Shift in accuracy (from Eq. 1) across several $\mathcal{D}$ and m values for example classifiers of different hypothetical abilities such that $\alpha \in [1,8]$.

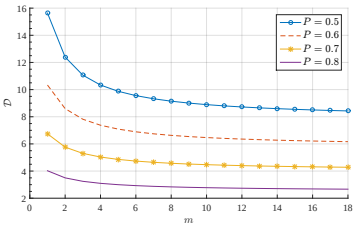

Fig. 4: Lower bounds on accuracy denoted by P that can be guaranteed with respect to task complexity D and the breadth of its solution space m for agent π^{mind} with ability $\alpha = 11.28$.

a minimal probability of success $P_{\mathcal{D},\alpha,m}$. For instance, if we know m, it is straightforward to calculate $\mathcal{D}$ from Eq. 1 as $e^{\frac{-\mathcal{D}}{\alpha}} = \frac{P_{\mathcal{D},\alpha,m} - \frac{1}{m}}{1-\frac{1}{m}} \Longrightarrow \mathcal{D} = -\alpha \ln\left(\frac{m \cdot P_{\mathcal{D},\alpha,m} - 1}{m-1}\right)$. Similarly a lower bound on accuracy can be guaranteed with respect to the task complexity and the breadth of its solution space. This is illustrated in Fig. 4 for agent π^{mind}.

This becomes interesting when a cost function (e.g. processing time, fee) is associated with utilising agents of higher abilities. Two agents π_1 and π_2, with abilities α_1 and α_2 and utilisation costs $c_1 = f(\alpha_1)$ and $c_2 = f(\alpha_2)$ respectively, guarantee an accuracy $P_{\mathcal{D}_1,\alpha_1,m} = P_{\mathcal{D}_2,\alpha_2,m}$ under different problem complexities such that $\mathcal{D}_2/\mathcal{D}_1 = \alpha_2/\alpha_1$. If $\alpha_2 > \alpha_1$ (and $c_2 > c_1$) then π_2 can accommodate (a α_2/α_1 factor of) higher problem difficulties with an additional cost of $c_2 - c_1$, while guaranteeing the same accuracy as π_1. Given a set of tasks of different complexities, a set of n agents of different utilisation costs, selecting the agent to solve these tasks with a *minimum bound on accuracy* of $\hat{p}$ can now be subsequently modelled as an optimisation problem: $\text{argmin}_{1 \leq i \leq n} f(\alpha_i)$, subject to $P_{\mathcal{D}_j,\alpha_i,m} \geq \hat{p}$.

Inferring task difficulty: alternatively, the IRT model can be applied to testing data in order to provide a quantitative understanding of the average complexity D of one class of tasks $X = \{x_1,...,x_t\}$, assuming the value m for such tasks is already known. For instance, one can empirically evaluate an agent of a known ability α over all task instances $x_i \in X$ and record its average score. Equation 1 can subsequently be solved for D using the recorded score as $P_{\mathcal{D},\alpha,m}$.

7 Collective Accuracy of Cooperative Agents

The advantages from adopting the IRT model extend to multiagent scenarios by estimating the collective accuracy of a group of agents. For instance, let A be a collective of agents using *simple majority voting* as a social choice function to elect a solution s_j to a problem x from the set of alternatives $S=\{s_1,s_2,...,s_j,...,s_m\}$ with only one correct solution $s_i \in S$. Let $Y=\{y_1,y_2,...,y_n\} \in S$, denote the votes of the agents in $A=\{\pi_1,\pi_2,...,\pi_n\}$ respectively regarding their preferred solution to x. When the votes are independent and identically distributed with equal accuracies p_x, the probability of collective A finding the solution s_j to x is:

$$P_x(A)=\sum_{k=\lfloor n/2 \rfloor+1}^{n} \binom{n}{k} p_x^k (1-p_x)^{n-k} \tag{2}$$

By combining equations 1 and 2, the probability $P_x(A)$ of a collective of agents $A=\{\pi_1,\pi_2,...,\pi_n\}$ electing the correct solution to x with difficulty $\mathcal{D}$, and alphabet m using simple majority voting becomes: $P_{\mathcal{D},m}(A)=\sum_{k=\lfloor n/2 \rfloor+1}^{n} \binom{n}{k} P_{\mathcal{D},\alpha,m}^k (1-P_{\mathcal{D},\alpha,m})^{n-k}$. This means that the probability of collective A solving a problem x is the sum of probabilities where at least 50% of the agents are correct. According to Condorcet's jury theorem [27], $P_{\mathcal{D},m}(A)$ is monotonically increasing when the IRT accuracy $P_{\mathcal{D},\alpha,m} > 0.5$ and vice versa. If A is a group of three agents with unequal accuracies of 0.55,0.55, and 0.63, its accuracy can be calculated from the agents' independent choices using majority voting as the probability of at least 2 out of 3 agents finding the correct solution: $(0.55^2 \times 0.37+2\times 0.45\times 0.55\times 0.63+0.55^2\times 0.63)=0.6144$. Similar predictions can also be performed using weighted[2] voting rules [16, Chap. 4]. The accuracy of an agent collective can thus be sometimes inferred from its agents' individual accuracies using the IRT model. Subsequently, one can analytically reason about the performance of groups of agents, in comparison to individual agent performance.

8 Analysing Individual and Group Accuracies

The accuracy of agent π^{mind} along with the accuracy of three agent collectives (A^1,A^2 and A^3) over different task complexities and solution spaces are illustrated in Fig. 5. We observe that adding agents of equivalent accuracies to the majority voting process (Collective A^1) improves the accuracy of the group over all tasks where the individual accuracy $P_{\mathcal{D},\alpha,m} > 0.5$, while the opposite is true for $P_{\mathcal{D},\alpha,m} < 0.5$. The key question here is, when is a (voting) collective more efficient than a single agent? To answer this, we calculate the *cut-off point* $\cap_{Y,Z}$ between two evaluated subjects Y and Z. To calculate $\cap_{\pi,A}$ (where the accuracy $P_{\mathcal{D},\alpha,m}$ of an agent π, and $P_{\mathcal{D},m}(A)$ of a collective A, are both equal over some task of complexity $\mathcal{D}$) we look for the value of $\mathcal{D}$ at which $P_{\mathcal{D},\alpha,m}=\frac{1}{m}+e^{\frac{-\mathcal{D}}{\alpha}}\left(1-\frac{1}{m}\right)=P_{\mathcal{D},m}(A)$, which leads to $\mathcal{D}=-\alpha\ln\left((P_{\mathcal{D},m}(A)-\frac{1}{m})/(1-\frac{1}{m})\right)$. If all the agents have similar accuracies (Collective A^1, Fig. 5), then according to Eq. 2, they are only equally accurate when $P_{\mathcal{D},\alpha,m}=P_{\mathcal{D},m}(A)=0.5$ leading to a $\mathcal{D}=-\alpha\ln\left((\frac{1}{2}-\frac{1}{m})/(1-\frac{1}{m})\right)=-\alpha\ln((m-2)(2m-2))$. For example, the cut-off point $\cap_{\pi^{mind},A^1}$ between π^{mind} with $\alpha=12.0094$ and A^1 over a problem with $m=3$ occurs at a $\mathcal{D}=-12\ln\left(\frac{1}{4}\right)=16.64$, which can also be verified from the graph in Fig. 5.

[2] More sophisticated voting rules such as *Borda count, harmonic rule, maximin* and *Copeland* require the subject to output a concrete ranking over all possible alternatives of the test/task, which inhibits our ability of making exact predictions. Yet, one can still analytically place $\min$ and $\max$ bounds on team accuracy using different sampling techniques.

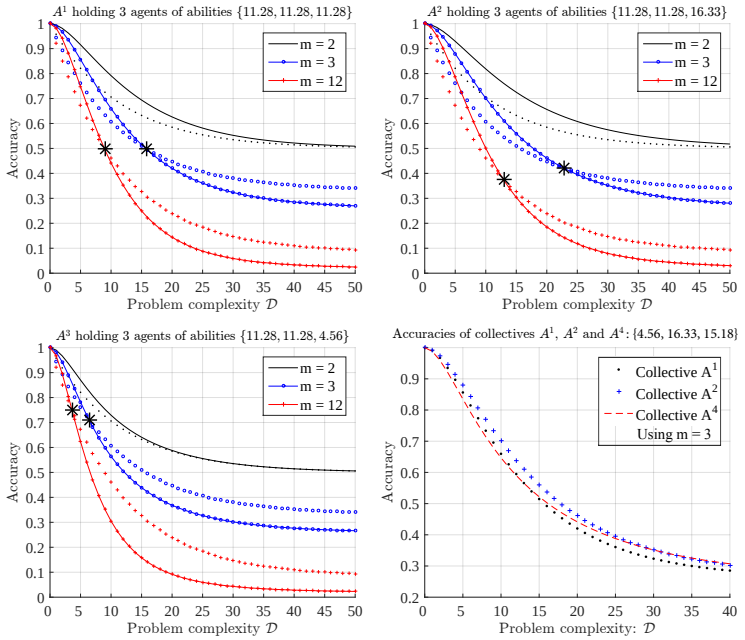

Fig. 5: Collectives accuracies aggregated using majority voting. The accuracy of π^{mind} is also depicted as dotted markers in the backgrounds of the first 3 plots for comparison. The $*$ symbol denotes the cut-off point where the accuracy of π^{mind} meets the corresponding group accuracy.

The cut-off point not only returns the setting over which $P_{\mathcal{D},\alpha,m}$ and $P_{\mathcal{D},m}(A^1)$ are equal, but also illustrates the relationship between the complexity of the problem $\mathcal{D}$ and the breadth of its solution space m, with respect to the accuracy of the evaluated group. In other words, the cut-off point indicates the problem complexities and solution spaces over which a collective is more effective than its individual agents. In most real world scenarios voting agents have different abilities and consequently different accuracies. Replacing a group member by another of higher/lower accuracy (Fig. 5 top-right/bottom-left) improves/diminishes the performance of the group by a measurable amount. For instance, let $A=\{\pi_1,\pi_2,\pi_3\}$ be the group of agents with abilities $\alpha_1,\alpha_2,\alpha_3$ and IRT accuracies (abridged as) p_1,p_2,p_3 respectively over some task x. If the agents' individual votes are independent, the probability $P_{D,m}(A)$ of A correctly guessing the solution to task x by majority voting is: $p_1p_2(1-p_3)+(1-p_1)p_2p_3+(1-p_2)p_1p_3+p_1p_2p_3$. When $p_1=p_2=p_3$, then $P_{D,m}(A)$ is equivalent to Eq. 2. If $A'=\{\pi_1,\pi_2,\pi'_3\}$ is the group of agents with accuracies p_1,p_2,p'_3 respectively s.t. $p'_3>p_3$, then its accuracy increases by $P_{\mathcal{D},m}(A')-P_{\mathcal{D},m}(A)=p_1p_2(p_3-p'_3)+(1-p_1)p_2(p'_3-p_3)+(1-p_2)p_1(p'_3-p_3)+p_1p_2(p'_3-p_3)=(1-p_1)p_2(p'_3-p_3)$ since $1/m\leq p_1,p_2\leq 1$ by definition (Eq. 1). For $p_1=p_2\neq p_3$ the cut-off point $\cap_{\pi_1,A}$ occurs at $\mathcal{D}=-\alpha_3\ln((p_3-\frac{1}{m})/(1-\frac{1}{m}))$ when $p_3=0.5$. As a result, we can measure the rise/drop in accuracies of A^2 and A^3 illustrated in Fig. 5 top-right/bottom-left. For example, for tasks of $m=3$, $\cap_{\pi^{mind},A^2}$ (Fig. 5 top-right) occurs at a $\mathcal{D}=-16.33\ln((0.5-\frac{1}{3})/(1-\frac{1}{3}))=22.64$.

Comparing agent collectives: scores from standard IQ tests provide us with some sort of scale or ranking of performances of evaluated individuals or groups. Nonetheless, these

performance measures might not be valid over certain settings. We observe in Fig. 5 that voting collective A^1 is more efficient than A^4 (holding agents with abilities $\{4.56,16.33,15.18\}$) over inference tasks of $\mathcal{D}<14$, whereas (counterintuitively) A^4 scores higher that A^1 over the C-test ($0.51>0.38$). Moreover, the opposite is true for tasks of higher complexities. Such scenarios might create confusions as they are frequently encountered and cannot be disclosed from standard intelligence tests. We also observe that for highly complex tasks with $\mathcal{D}>25$ collectives A^1 and A^2 record very similar accuracies since $P_{\mathcal{D},m}(A^1) - P_{\mathcal{D},m}(A^2)$ becomes very small. This is coherent with real world observations (although it cannot be drawn from intelligence test scores) as the accuracy of a subject, or a group of subjects, over extremely hard tasks is likely to converge to a random guess (an asymptotic minimum).

9 Conclusion and Future Work

IQ scores can be an unreliable predictor of an agent's performance over tasks of well-defined complexities and other problem settings. We proposed a new mathematical model that is flexible enough to predict the accuracy of agents of different abilities over various problem settings. We illustrated the relationships between the accuracy (and ability) of an agent, the complexities of the assessment task and the depth of its solution space. We identified agents that can guarantee a lower bound on accuracy with respect to task complexity and the breadth of its solution space. We further analysed settings over which a group of (majority voting) agents can be more or less effective than individual agents, or other groups. For instance, we directly inferred from the model the complexity at which a group is expected to record a similar accuracy as an individual agent, and beyond which a single agent is more effective than the group. We also measured the effect (on accuracy) of introducing agents of higher or lower abilities to a group of agents. Finally, we identified possible circumstances that are somewhat counterintuitive to the conclusions drawn from intelligence tests. These occur when a group of agents scores higher than another on an intelligence test yet fails to outperform this same group over some task complexities. In our future work more sophisticated voting rules will be used to analytically reason about team accuracy by analysing the outcomes from different sampling techniques over the agents' ranked votes.

References

1. Bien, Z., Bang, W.C., Kim, D.Y., Han, J.S.: Machine Intelligence Quotient: its measurements and applications. Fuzzy Sets and Systems 127(1), 3–16 (2002)
2. Birnbaum, A.: Some latent trait models and their use in inferring an examinee's ability. Statistical theories of mental test scores pp. 395–479 (1968)
3. Chmait, N.: Understanding and measuring collective intelligence across different cognitive systems: An information-theoretic approach (extended abstract). In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI-17 Doctoral Consortium. Melbourne, Australia (2017), (To appear)
4. Chmait, N., Dowe, D.L., Li, Y.F., Green, D.G., Insa-Cabrera, J.: Factors of collective intelligence: How smart are agent collectives? In: Proc. of 22nd European Conference on Artificial Intelligence ECAI. Frontiers in Artificial Intelligence and Applications, vol. 285, pp. 542–550. IOS Press (2016)
5. De Ayala, R.J.: The theory and practice of item response theory. Guilford Publications (2013)
6. Dowe, D.L., Hernández-Orallo, J.: IQ tests are not for machines, yet. Intelligence 40(2), 77–81 (2012)

7. Dowe, D.L., Hernández-Orallo, J., Das, P.K.: Compression and intelligence: Social environments and communication. In: Proc. of the 4th Int. Conf. on AGI. pp. 204–211. Springer, Berlin (2011)
8. Gottfredson, L.S.: Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. Intelligence 24(1), 13–23 (1997)
9. Hernández-Orallo, J.: Beyond the Turing Test. Journal of Logic, Language and Information 9(4), 447–466 (Oct 2000)
10. Hernández-Orallo, J.: The measure of all minds: evaluating natural and artificial intelligence. Cambridge University Press (2016)
11. Hernández-Orallo, J., Dowe, D.L.: Measuring universal intelligence: Towards an anytime intelligence test. Artificial Intelligence 174(18), 1508–1539 (Dec 2010)
12. Hernández-Orallo, J., Insa-Cabrera, J., Dowe, D.L., Hibbard, B.: Turing machines and recursive Turing tests. In: AISB/IACAP 2012 Symposium "Revisiting Turing and his Test". pp. 28–33 (2012)
13. Insa-Cabrera, J., Dowe, D.L., España-Cubillo, S., Hernández-Lloreda, M.V., Hernández-Orallo, J.: Comparing humans and AI agents. In: AGI'11. LNCS, vol. 6830, pp. 122–132. Springer (2011)
14. Klein, G.A., King, J.A.: A test for the performance of knowledge-based systems: Aiq. In: Proc. AAAI Workshop on Validation and Verification of Expert Syst. Menlo Park, CA (1988)
15. Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Problems of information transmission 1(1), 1–7 (1965)
16. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience (2004)
17. Legg, S., Hutter, M.: Universal intelligence: A definition of machine intelligence. Minds and Machines 17(4), 391–444 (Dec 2007)
18. Lempel, A., Ziv, J.: On the Complexity of Finite Sequences. Information Theory, IEEE Transactions on 22(1), 75–81 (Jan 1976)
19. Levin, L.A.: Universal sequential search problems. Problems of Information Transmission 9(3), 265–266 (1973)
20. Li, M., Vitányi, P.: An introduction to Kolmogorov complexity and its applications (3rd ed.). Springer-Verlag New York, Inc. (2008)
21. Lord, F.M., Novick, M.R.: Statistical theories of mental test scores. Addison-Wesley (1968)
22. Martínez-Plumed, F., Prudêncio, R.B., Martínez-Usó, A., Hernández-Orallo, J.: Making sense of item response theory in machine learning. In: Proc. of 22nd Europ. Conf. on Artificial Intelligence (ECAI). Frontiers in A. I. and Applications, vol. 285, pp. 1140–1148 (2016)
23. Raven, J.C., Court, J.H.: Raven's progressive matrices and vocabulary scales. Oxford Psychologists Press Oxford, UK (1998)
24. Roid, G.H.: Stanford-Binet intelligence scales. Riverside Publishing Itasca, IL (2003)
25. Sanghi, P., Dowe, D.L.: A computer program capable of passing I.Q. tests. In: Slezak, P. (ed.) Proc. 4th Int. Conf. on Cognitive Science (ICCS/ASCS-2003). pp. 570–575. Australia (July 2003)
26. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal, The 27(3), 379–423 (July 1948)
27. Shapley, L., Grofman, B.: Optimizing group judgmental accuracy in the presence of interdependencies. Public Choice 43(3), 329–343 (1984)
28. Solomonoff, R.J.: A preliminary report on a general theory of inductive inference (Report ZTB-138). Cambridge, MA: Zator Co 131 (1960)
29. Spearman, C.: "General Intelligence", objectively determined and measured. The American Journal of Psychology 15(2), 201–292 (1904)
30. Thurstone, L.L.: Primary mental abilities. (1938), chicago Press
31. Turing, A.M.: Computing machinery and intelligence. Mind 59, 433–460 (1950)
32. Wechsler, D.: Wechsler adult intelligence scale-fourth. San Antonio: Pearson (2008)