I.G.

José Hernández-Orallo
jorallo@dsic.upv.es
DSIC, Universitat Politècnica de València*

March 15th, 2018

Abstract

This report summarises the key ideas for a simple concept of I.G. and its explanatory value for several areas.[1]

*Also visiting the Leverhulme Centre for the Future of Intelligence, University of Cambridge.

[1]This report is a collection of notes for archival reasons and internal dissemination. If you have found it, especially beyond 2018, there must probably be more polished and developed versions and spin-off papers from this one. Check the author's publication site for this.

Contents

1 Introduction

For more than a century, the intuitive idea of general intelligence has been associated with competence for a wide range of cognitive tasks. However, this interpretation has a critical pitfall. With limited resources, one has to choose which tasks to prioritise. A resource-bounded intelligent agent, be it an animal or an AI system, must *concentrate* its resources (brain, energy, computation, etc.) for some pockets of problems. As a result, this preference (or specialisation) for some problems over others would entail that a general intelligent system with limited resources could not even exist.

Still, for more than a century, there is sustained evidence for something that could be reasonably called general intelligence in humans [110, 73, 15], and other animals [3, 10]. Also, there has been a fundamental interest in AI to build general intelligence, from the early general problem solver [96], McCarthy's call for "generality in AI" [91], to the new expectations put on Artificial General Intelligence [1]. How can we reconcile this evidence with the intuition that resources must be prioritised for a limited pocket of tasks or environments?

In this paper, we disentangle this conundrum with the introduction of a simple notion of generality that is consistent with existing evidence in all these disciplines. This measure of generality is shown to be related to populational notions of general intelligence in humans, groups and non-human animals, the notion of cognitive efficiency and convergence in animal evolution, and the computational views of general intelligence based on Solomonoff priors and universal search.

Disclaimer: this report runs through very different disciplines around the notions of generality in cognition and general intelligence. This report is not meant to be comprehensive in the coverage of the literature. It introduces a new generality score and goes as directly as possible to those several issues in these disciplines that can be explained, predicted or better addressed with this notion. For a full coverage of the literature in all these disciplines, and especially the connections between them, the reader is recommended to have a look at [41].

2 General intelligence

The first scientific notion of general intelligence for humans, still prevalent today, was introduced by Charles E. Spearman [110]. He observed that humans who performed well (respectively poorly) for some cognitive tests usually performed well (respectively poorly) at the others. He made his observations from response matrices combining the results of respondents (humans) and tasks (test items), such as the one shown in Table 1 (left). From the results he derived the notion of the g factor, a latent factor that explained the variability in the matrix. Clearly, this concept of the g factor is populational, i.e., for some other sets of tasks and humans the g factor might be higher or lower. Consequently, the g factor does not provide a measure of the *generality* of an agent. Of course, if we are interested in the general intelligence of single persons, we could use

the g score, an individual value estimating how much of that latent variable each subject has, assuming some distribution parameters. However, a higher g score means that the person has more general intelligence – if we *reify* this concept from g –, but it does not indicate whether the person achieves the score by succeeding in many tasks but failing systematically with a specific subset of tasks. For instance, Table 1 (right) shows two agents that might have similar aggregate score but one (π_a) seems intuitively more general than the other (π_b).

Table 1: Left: Representation of a generic response matrix $r_{i,j}$ with columns representing the tasks (also referred to as tests or items) and the rows representing the agents (also referred to as persons or respondents). Correlations are usually analysed between tasks (columnwise). Right: Individually, we can look at the variances. Having the same average result (0.8), which agent, π_a or π_b, is more general?

	μ_1	$\cdots$	μ_N
π_1	$r_{1,1}$	$\cdots$	$r_{1,N}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
π_M	$r_{M,1}$	$\cdots$	$r_{M,N}$

	μ_1	μ_2	μ_3	μ_4	μ_5
π_a	0.85	0.75	0.80	0.85	0.75
π_b	1.00	1.00	0.00	1.00	1.00

In animal cognition research, especially in evolutionary terms, general intelligence is usually understood populationally as well (with intra- and inter-species factors, known as g and G). However, the drives for general intelligence are usually justified by how cognitively-demanding niches are. It seems reasonable that nervous systems evolve with resource constraints and it seems more beneficial to succeed in two scenarios A and B that have low cognitive demands than to succeed in one scenario C with higher cognitive demands. Given the same resources or evolutionary effort, a species that succeeds with A and B (but not C) is more expectable than a species that succeeds in C (but none of A and B), assuming the three scenarios are equally likely in the species's environment.

Another different view of generality is based on the *transitivity* of performance, which is an important indicator in some AI competitions. Actually, for the general videogame AI (GVGAI) competition [97, 9], it has been found that "performance is non-transitive", meaning that "different algorithms perform best on different games"[2]. Transitivity can be expressed as follows: if π_a is worse than π_b and π_a solves μ then π_b should also solve μ. That would suggest that π_b *dominates* π_a, and would make π_a redundant. Actually, if there is a strong correlation between tasks (and hence a high g factor), this is likely to happen. But again, this property (which by definition involves two agents) is understood for a *population* of agents.

[2]Even if the competition aims at *general* videogame playing, hence the name, the focus is ion finding non-transitivity, such that metalearning through hyper-heuristics and algorithm portfolios is effective [9, 94], by combining where some agents are good while others are bad, and vice versa. This is also a common thing in ensemble methods [78], where diversity is positive if results are to be combined.

In order to identify an *individual* measure of generality we could look at the variance of results for a particular individual. A low variance would indicate generality, as we can see in Table 1 (right). However, this is somehow assuming that all tasks have similar difficulty because, otherwise, we would expect agents to behave better for easy tasks and worse for difficult tasks. When selecting a battery of tests, the items are grouped and their difficulties are chosen so that we have tests that are informative. For instance, Spearman would have never considered a mathematics test for which all students score perfectly (all questions are easy) or all students fail systematically (all questions are difficult). However, in an uncontrolled and ungrouped scenario, tasks may be of a wide range of difficulties. Consequently, small variances would be simply unnatural.

The panorama changes completely if we consider *generality as solving a wide range of tasks up to a given difficulty*, a value that can also be roughly identified with the *capability* (and the internal capacity) of the individual. This correspondence – and duality – between difficulty and ability is at the core of cognitive measurement, especially in item response theory [23, 13], where latent factors are estimated from a population of items and respondents after the assumption of some parametric models and distributions.

In the next section we will introduce an individual (non-populational) notion of generality that takes difficulty into account in the first place. The rest of the paper explores its properties and the explanatory value of this new unifying notion for human intelligence (psychometrics), animal intelligence (comparative cognition) and machine intelligence (artificial intelligence).

3 Individual generality

We will consider the evaluation of a set of M agents on a set of N tasks, with results or responses $r_{i,j}$ for each agent π_j and task μ_i, as represented in Table 1 (left). For each agent we have its response mean $\bar{r}_j \triangleq \text{Mean}_i[r_{i,j}]$, also referred to as *agent average performance*, and its response variance $\sigma_j^2 \triangleq \text{Var}_i[r_{i,j}]$, also referred to as *agent variance*, defined as its *populational* variance[3].

We could simply define one notion of regularity as the reciprocal (inverse) of the variance. This would give us $1/\sigma_a^2 = 1/0.002 = 500$ and $/\sigma_b^2 = 1/0.16 = 6.25$ for agents a and b in Table 1 (right). But is the variance produced by unreliability in the measurement, instability in the agent or is it because the agent really performs much better at some problems than others? Let us first exclude all sources of unreliability and work at the definitional level. In order to do this, instead of actual responses, we are interested in expected (or ideal) responses. For each agent π_j and an instance or task μ_i, the expected response is given by $\mathbb{E}[r]_{i,j}$. We assume $0 \leqslant \mathbb{E}[r]_{i,j} \leqslant 1$ with 0 meaning worst possible performance and 1 meaning best possible performance. We now discretise expected responses as $A_{i,j} = 1$ ('acceptable' or 'accomplished') if $\mathbb{E}[r]_{i,j} \geqslant 1 - \epsilon$ and 0

[3]For binary responses, we have a Bernouilli distribution, and the variance is just reduced to $\bar{r_j} \cdot (1 - \bar{r_j})$.

otherwise ('unacceptable' or non-accomplished), where $1-\epsilon$ is just a threshold[4]. For instance, for dichotomous tasks (where agents can only be right or wrong), with an $\epsilon = 0.3$, we have that $A_{i,j}$ is 1 if the agent is expected to be correct on the instance at least 70% of the times.

This simple transformation eliminates reliability issues in our analysis of generality. But still, can we define generality as being good for all possible problems? First, for many sets of tasks N it is not possible to have acceptable results for all of them, as some may be very complex or may require more resources than the agent has. Second, if we use binary acceptability, we would have a Bernouilli distribution, and the variance would be linked to the average of results $(\bar{r_j} \cdot (1 - \bar{r_j}))$.

The way-out of these two problems is to look at responses in terms of their *difficulty*. Actually, agents might be better for easy problems than for hard ones. Any meaningful notion of generality should not ignore this possibility. Actually, it should place its quantification at its core.

3.1 Agent characteristic curves (ACCs) and capability

Let us then consider a function of difficulty[5], $\hbar$, mapping each task μ_i to a real value $\hbar(\mu_i) \geqslant 0$. We define an agent characteristic plot for agent π_j as a scatter plot showing accomplishment $A_{i,j}$ in terms of the difficulty $\hbar_i$. In other words, we plot difficulty on the x-axis and accomplishment on the y-axis.

We can convert these scatter plots (as the dots are always 0s and 1s) into more interpretable curves. In order to do this, we define $\psi_j^{[h]} \triangleq \mathbb{P}(A_{i,j} = 1|\hbar(\mu_i) = h)$, or equivalently, the mean of the accomplishment of agent j on all problems of difficulty h. We then define an agent characteristic curve (ACC) as a plot of $\psi_j^{[h]}$ as a function of h. Figure 1 shows six scatter plots (grey circles) and their corresponding ACC (blue line).

We say that an agent characteristic curve is s-saturated if $\forall h \leqslant s : \psi_j^{[h]} = 1$. We see that the two first ACCs are not even saturated for $s = 0$. We want agent characteristic curves to ensure that the area under these curves is finite. We will assume difficulty functions that meet this property. We will come back on this later but setting a threshold on tasks ensures this when difficulty is defined in terms of minimal resources [41], especially in situations where there is a minimum percentage given by chance. Another simpler option is just to set a maximum difficulty.

Capability is defined as:

$$\psi_j \triangleq \int_0^\infty \psi_j^{[h]} \, dh \tag{1}$$

i.e., the sum of all the mean responses per difficulty, which is the area under the ACC (see Figure 1 for the calculated capabilities). Note that in a discrete way,

[4]The value of ϵ might be different for each task. Actually, by changing the threshold we change the difficulty of the task, which is actually like having another task.

[5]As we will see in the following sections, the difficulty function can be derived from the characteristics of the tasks or it can be derived experimentally from the results of a population.

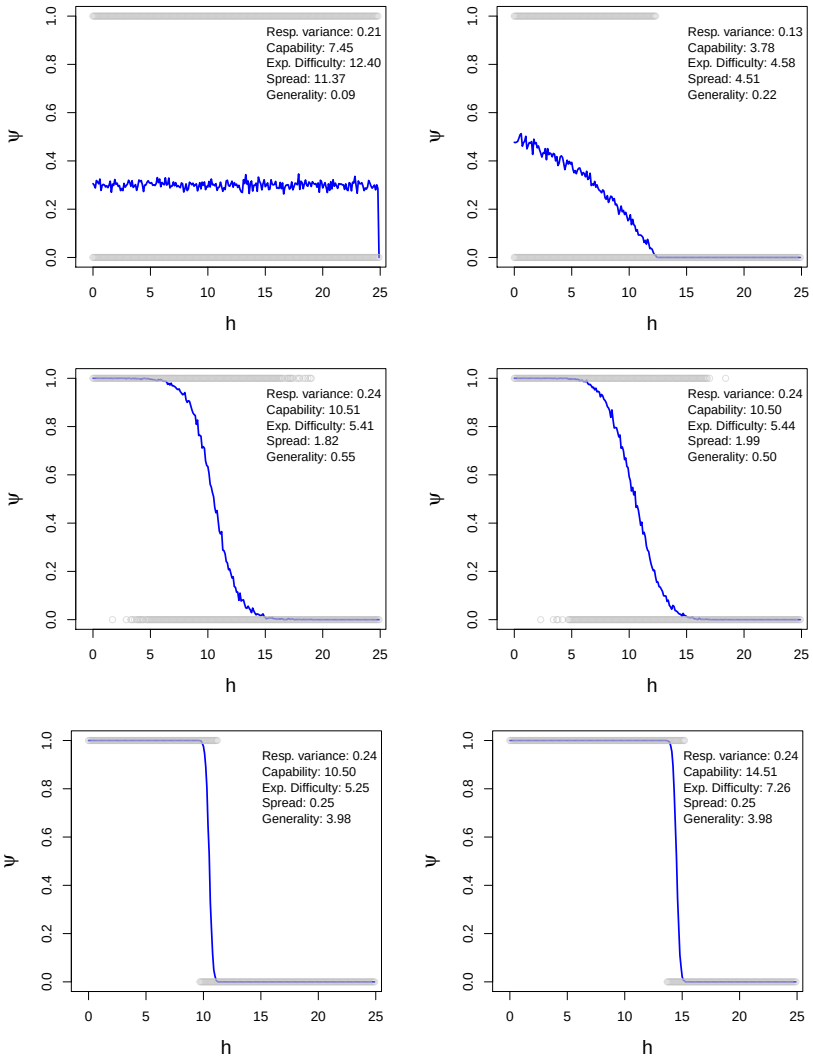

Figure 1: *Agent characteristic curves* (ACC), showing the behaviour of six different agents in terms of difficulty $\hbar$ on the x-axis. The responses $r_{i,j}$ for the items i are shown in grey circles. The means for each difficulty are shown in blue, and connected to form an ACC. We see that different distributions of results give different values for the metrics: response variance (σ_j^2), capability (ψ_j), expected difficulty ($\mathbb{H}_j$), spread (z_j) and generality (γ_j). Curves that have a steplike shape have high generality.

capability is a weighted sum of all tasks according to a prior uniform distribution (or weight) of difficulties. The area will of course change even with a monotonic transformation on difficulty, such as a change to a logarithmic scale. Some scales make more sense than others and give a more meaningful notion of capability, especially if the x-axis can be associated with an additive unit, as we will discuss later on. These weights can be derived if we know the posterior, how many tasks we have for each difficulty.

3.2 Definition of individual generality

In order to introduce a measure of generality that accounts for agents that perform generally well for all problems *of low difficulty*, we must look at how *compacted* the curve is on the left, or how much step-like it is. This is tantamount to analysing how well-employed the *effort* to achieve the capability is, understanding that this effort is given by the difficulty of the task. This is actually the first partial moment of the ACC for agent j, denoted by m_j, and calculated as follows:

$$m_j \triangleq \int_0^\infty h \cdot \psi_j^{[h]} \, dh \tag{2}$$

As capability represents the "mass" (how much of accomplishment we have), we can normalise this moment (dividing by capability, so that $\psi_j^{[h]}$ is normalised to a density function), and we have the expected difficulty for agent j:

$$\mathbb{H}_j \triangleq \mathbb{E}_i[h|A_{i,j} = 1] = \frac{m_j}{\psi_j} \tag{3}$$

In other words, $\mathbb{H}_j$ is actually the expected difficulty if we sample items using the agent accomplishment as probability.

Now, if we look at $\mathbb{H}_j$ as an expected difficulty, then, for a distribution that is fully compacted on the left (a step function), this should be half of the capability. This difference (multiplied by capability back again and finally squared rooted, to make it independently of location and with a unit commensurate with difficulty, as we will see) is known as spread, and is given by:

$$z_j \triangleq \sqrt{(2\mathbb{H}_j - \psi_j) \cdot \psi_j} = \sqrt{2m_j - \psi_j^2} \tag{4}$$

If we take capability as a location on the x-axis, spread would be interpreted as measuring how much (and how far) the mass spreads over the left and right of that location.

Finally, we just define generality as the reciprocal of spread, i.e.:

$$\gamma_j \triangleq \frac{1}{z_j} = \frac{1}{\sqrt{2m_j - \psi_j^2}} \tag{5}$$

It is important to note that γ_j is just a metric that can be applied to any possible set of points or a function in the domain $\geqslant 0$ and in the range $[0, 1]$, provided

the used difficulty leads to a finite area under the curve. There is no parametric model assumed. Actually, γ_j does not rely on any model. For instance, the middle left plot in Figure 1 follows a logistic function while the middle right and the two bottom plots in Figure 1 follow an error function (proportional to the cdf of a normal distribution).

3.3 Properties of generality

Before looking into how the definition of generality fits in with several disciplines, let us analyse its formal properties, jointly with capability. First, we need to introduce the notion of difficulty *translation*, defined as a constant shift on the x-axis ($h + k \leftarrow h$). If k is negative we have a translation to the left, where every result with $h < 0$ is cut out. If k is positive we have a translation to the right, and we assume that $\psi_j^{[h]} = 1$ for all $h < k$ (i.e., we saturate the newly introduced part of the curve). Second, we introduce the notation for partial areas, i.e., $\psi_j^{[h_1:h_2]} \triangleq \int_{h_1}^{h_2} \psi_j^{[h]}\, dh$.

Now we have some simple properties[6]:

1. Translation: any positive translation by k implies that capability becomes $\psi_j + k$. The same happens for negative translation if the $|k|$-leftmost part of the original curve was saturated. On the other hand, generality is invariant to translation (with the same conditions as above for negative translation).

2. Compactness: with equal capability, any equal mass moved to the left of the plot such that $\psi_j^{[h_1:h_2]} \leftarrow \psi_j^{[h_1:h_2]} + q$ while $\psi_j^{[h_3:h_4]} \leftarrow \psi_j^{[h_3:h_4]} - q$, with $h_2 < h_3$, will increase γ_j.

3. Maximum: given a fixed capability ψ_j, the minimum expected difficulty $\mathbb{H}_j$ and the maximum generality γ_j are obtained with a step agent characteristic function on $h = \psi_j$, where the capability is double the expected difficulty (i.e., $\psi_j = 2\mathbb{H}_j$), and generality $\gamma_j = \infty$.

4. Given a constant function $\psi_j^{[h]} = c$ from 0 to q, we have $\psi_j = cq$, $z_j = \sqrt{c(1-c)}q$ and $\gamma_j = 1/\sqrt{c(1-c)}q$. In the particular case of $c = 0.5$ we have $z = q/2$ and $\gamma = 2/q$.

5. Task transitivity: if an agent π_j is s-saturated then for every task μ_b such that $A_b^j = 1$ in the saturated area then fo all other tasks a of $\hbar(\mu_a) \leqslant \hbar(\mu_b)$ we have that $A_a^j = 1$. In other words, if this agent solves a task in the saturated area then it also solves any other easier task. Agents with maximum generality $\gamma_j = \infty$ are s-saturated with $s = \psi_j$, so once a task of a given difficulty is solved there is no need of checking easier tasks.

[6]The proofs are straightforward, but they can be found in the appendix.

6. Agent transitivity: if two agents π_a and π_b have maximum generality $\gamma_a = \gamma_b = \infty$ and $\psi_a \leqslant \psi_b$ then for every task μ_i such that $A_i^a = 1$ we have that $A_i^b = 1$. That means that π_b *dominates* π_a or, in other words, that an agent would solve all tasks a less capable agent solves, provided both have maximum generality. Note that if generality is not infinite, it is not sufficient to have a curve for π_b that covers the curve for π_a. We need to check that π_b is s-saturated for at least the maximum value where π_b gets non-zero accomplishment.

7. Same units: if we introduce a unit for difficulty, let us call it *bints* (for basic intelligence units), then capability is also (additively) measured in *bints*, spread is also measured in *bints* and hence generality is measured in $1/bints$.

Some of these properties (especially the transitivities) have been shown when assuming a Guttman (or deterministic) response model [30, 31], as we will discuss in the next section. Looking again at Figure 1 we see that the bottom left and bottom right are basically a translation of each other by $k = 4$. We see that the capability is increased by approximately 4, and the spread and generality are not significantly affected.

The translation property (#1) shows a way of increasing capability without losing generality, an indication of how more general and more capable systems are possible, for both artificial and natural systems. In principle, capable systems do not have to be necessarily more or less general, which links well with the question of Spearman's law of diminishing returns, as we will explore in the next section. But note that the compactness and maximum properties (#2 and #3) suggest an optimal way of getting a given capacity, if the difficulty of a task is understood as (or related to) the effort for finding a solution to the task. We will revisit this problem in relation to the convergent evolution of intelligence and theoretical measures of computational effort and difficultly. Finally, the task and agent transitivities relate to many issues of measurement in human intelligence and artificial intelligence, from the g factor to the c factor.

In the following sections we will explore some and other of these phenomena about general intelligence in three major disciplines: human intelligence, mostly in psychometrics, animal intelligence from an evolutionary perspective and AI, especially about the quest of AGI and metrics of general intelligence. The following sections are meant to be sufficiently independent for those readers that are interested in or familiar with only one or two of these disciplines.

4 Psychometric interpretation: generality, the *g* factor, SLODR and the *c* factor

In this section we will analyse the interpretation of the notion of generality in the context of the science and literature of human intelligence. We will first flesh out the clear connections and inspirations, and then we will explore some other more profound implications.

4.1 Related metrics and models: person-fit, Guttman scales, reliability and variable-θ models

Psychometricians will find the previous sections familiar in some ways. The use of two parameters, difficulty for items, and ability for subjects is common in classical test theory and especially in item response theory [23, 13]. Also, plotting the performance, or the probability of correct response, against ability on the x-axis leads to the item characteristic curves. Similarly, plotting this against difficulty on the x-axis leads to subject or person response curves [117, 115]. It is important to note, however, that in IRT, both ability (usually denoted by θ) and difficulty (usually denoted by b in logistic models) are latent factors, which are estimated by making several assumptions: "1) local independence, 2) unidimensionality, and 3) a specified shape for the item characteristic curve." [114]. The shape is determined by a model, which is usually a decreasing monotonic function on $b-\theta$, such as a logistic function. Then the parameters are estimated from a response matrix $r_{i,j}$.

In our case, we are not considering a measurement problem (yet), and we are not (necessarily) plotting latent variables. Difficulty could be a notion derived from the items themselves, and capability —and not ability— is not the parameter of any function. Actually, we define capability —and we use a different term on purpose— as an area, and not the location of the steepest point of any curve. For models that are symmetric at $y = 0.5$, such as 1PL or 2PL logistic models, the area equals this location. However, for irregular curves not following a model at all, it is the area what is really meaningful. Also, we are not plotting correct response, but whether the *expected* response is above a threshold, because we are considering the true/expected values (not the measurement problem where we only have one or a few samples of each pair of item and person).

The key question about the assumptions in IRT is that even if some models allow for a discrimination parameter for the items, so that that the correlation between correct response and ability for all items is relaxed (it might even be negative), this is not usually the case for the ability. For many models, IRT is actually assuming a strong (negative) correlation between correct response and difficulty for all agents. Note, by the way, that a step ACC does not maximise (negative) correlation (it is actually -0.866). The models (not even the variable-θ ability models we will mention below) consider that a subject being better at difficult items than easy items is an aberration, mostly because the models and estimations are done in such a way that this is assumed not to happen (or should just show a bad fit to the model).

This has actually led to a myriad of person-fit metrics [93], which is a way of analysing subjects at the individual level. This aims at identifying cases such as "low-ability examinees who copy answers to several difficult items from a much more able neighbor and very high-ability examinees fluent in another language but not yet fluent in English, who misunderstand the wording of several relatively easy questions" [114]. But in the end, all this is about whether the observed curve matches the expected curve. This was not meant to measure generality. Actually, a step function is usually categorised as overfitted by

person-fit metrics, because it deviates from the usual logistic models (for which infinite discriminations and hence infinite slopes are rare).

As there are so many person-fit metrics, some of them are relatively similar to γ, as we defined in previous sections. Especially relevant are those that compare the person response curve with a Guttman conformal curve, which is a curve that is right for the first r items of lowest difficulty and wrong for the rest (a step function). In this setting, the closest metric seems to be the *norm conformity index* [113], which basically counts how many ranking mismatches there are between a Guttman curve and an observed curve. Another very related metric is the disagreement index [75], where the agreement index (the sum of the results multiplied by the difficulty index for all items) is compared with the score of the Guttman conformal curve with the same *number of correct* responses (NC score). Since all these metrics are ordinal, and convert the difficulty of the items to ranks (index), the correspondence to γ is only direct when we have a uniform distribution of items per difficulty. In other words, all these metrics take all instances as equally valuable —the NC score is the number of counts, the number of correct responses—, while the agent characteristic curves shown in Figure 1 sum with the assumption of difficulties being uniformly distributed. So, if there are more items for some difficulty values than others, the count (the NC score) and the area (the capability) would be different and all metrics would differ. This is intentional, as we are not interested in a capability according to a set of items, but according to different levels of difficulty. We are at the theoretical level, or on expectation. Actually, for many difficulties, the number of items might be infinite. So, assuming an uneven number of items per difficulty does not have more support than assuming them uniform.

Still, because many of these metrics take the step function as a reference, it is important to look at the Guttman scale or, more precisely, the deterministic model [30, 31], which can be considered a precursor of IRT. A deterministic model just captures the item response curve as a step, i.e., the probability of correct response is 0 for values below the ability θ and elsewhere. This model produces agent response curves that are also a step —the Guttman conformal curves— and, hence, they would have infinite generality. Several properties derive when items (and hence agents) follow this model. In particular, task transitivity and agent transitivity are true under this model, as we have shown in section 3.3 (properties #5 and #6).

The Guttman scale assumes monotonicity (higher probability of response for higher ability), but there are many other models (some non-parametric [95, 105] and some parametric [23, 13]) assuming this. The Guttman model has been used in cases where solving one item means all items of lower levels of difficulty have to be solved as well. For instance, at the lowest level of difficulty one might have addition and then at the next level we can have multiplication. Arguably, one cannot do any multiplications without knowing addition (although there are very simple cases such as multiplications by zero or by one that do not require any addition in the process). In general, the Guttman model does not hold for practical sets of items, and it is mostly used because of its simplicity.

It is important, hence, to say that our notion of generality is not assuming

the Guttman model for items (or a non-ordinal version of it) or a conformal Guttman curve, but just measuring how far the expected responses of an agent are from that theoretical situation.

Finally, there is a clear resemblance of the notion of generality with "person reliability", as introduced by Lumsden [85]. The notion of reliability wants to capture "tremor effects", i.e., each person has a variability on its ability θ. Actually, Lumsden models this reliability with a normal distribution and then the agent characteristic curve turns out to be its CDF. For constant-θ IRT models, like the traditional logistic models or the Guttman model, the theoretical agent characteristic curve has the same slope for all respondents. This changes for variable-θ IRT models, where reliability is introduced as an extra parameter (sometimes sacrificing the discrimination parameter, depending on the degrees of freedom).

In general, without considering any particular model, an agent can get constant θ, with no reliability issues at all, and still have a flat curve. Simply, the agent is consistently bad at easy problems, like the two top plots on Figure 1. It is only when we limit ourselves to some particular models that we can understand the slope of the curve as a reliability. In other words, variable-θ models assume "that the person trait level varies during test administration" [25]. By using expected values and thresholds transforming them into accomplishment values we exclude the reliability component and we focus exclusively on generality.

Perhaps because of this confusion between reliability and generality, the agent reliability metrics are not as widespread as the person-fit metrics commonly used for constant-θ IRT models. But we have to be careful about person-fit: "From a constant-θ point of view, person reliability can be considered as a source of misfit or overfit at the individual level. Thus, the imprecise, highly unreliable respondent [...] will produce an almost random pattern that will be regarded as misfitting. At the other extreme, the highly reliable respondent is expected to produce a highly scalable response pattern that fits the stochastic model too well and that will be regarded as overfitting". Here, in contrast, with the individual metric of generality, we are not considering any model to fit. For generality we just examine the distribution of the expected responses in terms of difficulty.

Once the differences between generality and reliability are clarified at the conceptual level, we may be interested in the connections at the formulaic level. For instance, if we generate expected responses according to a normal distribution (like the middle right and the two bottom plots in Figure 1), with a standard deviation σ we have the following[7]:

Proposition 1. *Assuming a normal distribution on the capability, with standard deviation σ, the slope of the ACC will be $-\frac{1}{\sigma\sqrt{2\pi}}$.*

Less trivially, we can show the following lemma and proposition:

[7] Proofs in the appendix.

Lemma 2. *Assuming a normal distribution on the capability, with mean μ and standard deviation σ, such that the location is sufficiently beyond 0 to have negligible mass below 0 (i.e., $\frac{\mu}{\sigma} \gg 0$), we have that $m_j = \frac{\sigma^2+\mu^2}{2}$.*

Proposition 3. *With the same assumptions as lemma 2, we have that spread $z_j = \sigma$ and $\gamma = \frac{1}{\sigma}$.*

The definition of person reliability was just $1/\sigma$ [25], so we see the equivalence between reliability and generality if the agent had an ACC that were complementary of a normal CDF. However, in our case we do not understand σ as the standard deviation of capability or its measurement and there is no special reason why this should be normal.

4.2 From individual generality to populational generality: manifolds and the *g* factor

As we mentioned in section 2, Charles Spearman found an important phenomenon; when he analysed a set of different tests taken by the same population, and calculated the correlations between tasks ($\rho_{a,b}$ for each pair of tasks a and b on a result matrix such as the one shown in Table 1), he found a positive average correlation ($\bar{\rho} \gg 0$). In other words, the individuals that obtained good results for some tests usually obtained good results for the rest. This correlation was stronger the more culture-fair and abstract the tests were. This phenomenon was known as the 'positive manifold' [110, 111]. It is important to clarify that this phenomenon is not a property of the tests (tasks) alone nor a property of the population (agents) alone. A correlation is clearly an effect that takes place for two tests for a set of subjects, but the average correlation is calculated from the correlation matrix, thereby involving all the tests and all the agents in the population. Nevertheless, the positive manifold appeared again and again for different human populations and different sets of tests, provided they were not too linked to particular cultural or educational backgrounds (e.g., a chess-playing test or a Korean vocabulary test).

Spearman tried to understand the findings through the invention of a rudimentary factor analysis. He identified a dominant *latent factor* that explained much of the subjects' variance, and called it the *g* factor. Since then, this factor has been one of the most relevant (and replicated) findings in psychometrics [73, 112] and has been found to predict many facets of life, from academic performance to (lack of) religiosity in humans. The dominance of *g* and its explanatory character for the positive manifold led to the association of *g* with general intelligence, a latent factor that was said to pervade all other factors and facets of intelligence. Of course, this interpretation has been challenged many times, even if *g* appears again and again.

Note that the theory behind *g* allows psychometricians to estimate this factor for individuals, giving us a latent factor that is useful in general. But still, ability and generality are two different things. For two different people with the same *g* score, we could have that one person achieves good results for other cognitive

tests systematically but another person may get a more uneven performance for the same set. In other words, the predictability of g scores is analysed globally, but still some individuals may be less predictable than others. One possible reason may be reliability[8], but another reason is simply that some individuals are less general than others. So the question was whether a general factor emerges from human performance on a range of tests. But where does this general factor come from in the first place? Is it a necessary result if the individuals are general? This new question is what we try to explore below.

Let us first analyse the situation where all agents have maximum generality. Without loss of generality, we can consider that the rows of the response matrix $r_{i,j}$ are ordered by increasing capability. This means that all columns in the response matrix $r_{i,j}$ would be of the following form 1^p0^q, with $p + q = M$, i.e., the item response curves would follow a Guttman model. If $p > 0$ and $q > 0$ the correlations will be well defined and will be strictly greater than 0 and there will be a positive manifold. Depending on the distributions of capabilities and difficulties the magnitude of the average correlations will vary. For instance, it is easy to see that if we consider a normal distribution of difficulties and an equal normal distribution of capabilities, the mean correlations will be around 0.47, which is a very important positive manifold. In this situation, we see that individual generality implies a positive manifold. We do not even need a factor analysis to say that the individual generality extends as a populational generality. As Guttman points out, a notion of *populational* generality can just be defined "as having all correlations positive or zero", without the need of "a common factor" [32].

Spearman, and most of the literature after him, analysed the positive manifold for tests instead of items. Tests group a number of items that are considered to be related (e.g., a maths test) and include a range of difficulties so that we get a range of results for the test according to the population it is going to be applied to. So let us consider that items or tasks μ_i are grouped into tests τ_k. Now we can construct a new response matrix where columns k are tests and rows are agents j. We can analyse that by aggregating items into tests, mean correlations may get much higher under different scenarios.

For instance, let us consider both item difficulties and capabilities following the same normal distribution (sufficiently far from 0 so that there is negligible mass below 0). In this case, we have that if we group the items randomly, we can get mean Pearson correlations above 0.99. In general, we can relax a sample grouping as long as the new groups preserve the item difficulty distributions. With this condition, for each agent π_j we will have exactly the same results for all tests. As the agents have different capabilities, we will have a mean Spearman correlation equal to 1 and, if the distributions are normal, a very high mean Pearson correlation.

Other similar results can be obtained with some other distributions, assuming that each test preserves a range of difficulties such that ensures the

[8] With *g=1* we can still have that each agent fail a different percentage of the times, but in a completely random way. Actually, by taking a perfect agent and introducing different levels of systematic noise to form a population, one would get perfect *g*. This is not generality.

differences in capabilities to be represented per each test. This is actually a very natural condition for a test to be informative. If all respondents got similar values for the test, then the test would not very informative. As a result, the only strong sufficient condition for a high manifold to appear is individual generality.

Note that we have seen that individual generality implies populational generality (sufficiency), but this does not exclude that populational generality could have been obtained by other means, with all agents with different abilities but flat ACCs, as the one on Figure 1 (top left). This situation would actually require fewer conditions on the distribution of difficulties (actually difficulties would not play a role for these curves up to the point where the flat curve stops) but will necessarily require a random sample per difficulty (one could even get negative manifolds if instances are chosen on purpose to do so). Ultimately, dominance between ACCs (not only different capabilities) would be a more refined condition for this.

But only the maximally general ACCs can ensure that for every possible partition or sampling of instances, provided the range/distribution of difficulties is kept, the manifold is created, since the capabilities are preserved for each subtest. This is illustrated at the top of Figure 2 in contrast with the bottom of the same figure.

In sum, if individuals have low generality, choosing sets where a difficulty range is preserved is less important, and the positive manifold could still appear if the tests are not splitting the items by pockets of speciality. If this is not the case, the manifold might even be negative. On the other hand, if individuals have high generality, any partition of items into tests provided the range of difficulties is preserved would lead to high positive manifolds. In any case, negative manifolds would never appear.

Negative manifolds are very rare in the literature of human intelligence. Also, having sets of cognitive items for which difficulty does not play a role seems very unnatural. But still, the evidence might be compatible with some moderate degrees of generality or some individuals being more general than others. The plausibility (or necessity) of a particular scenario in light of a positive manifold will depend on a series of assumptions. Of course, the sufficiency direction is clear: if we are able to measure generality of the individuals in a population and we know how tests are formed, we can predict the manifold.

4.3 Spearman's Law of Diminishing Returns (SLODR) and individual generality

There is another source of evidence that can help us with the analysis of the plausibility of individual generality in light of a positive manifold. This evidence was also first gathered by Spearman. He calculated the strength of g on subpopulations of different abilities. In particular, in one of the analysis, he separated the results of several tests on a human population into two groups: group A with normal abilities and group B with low abilities. After the split, he analysed the correlation matrices separately. The result was that the mean correlation

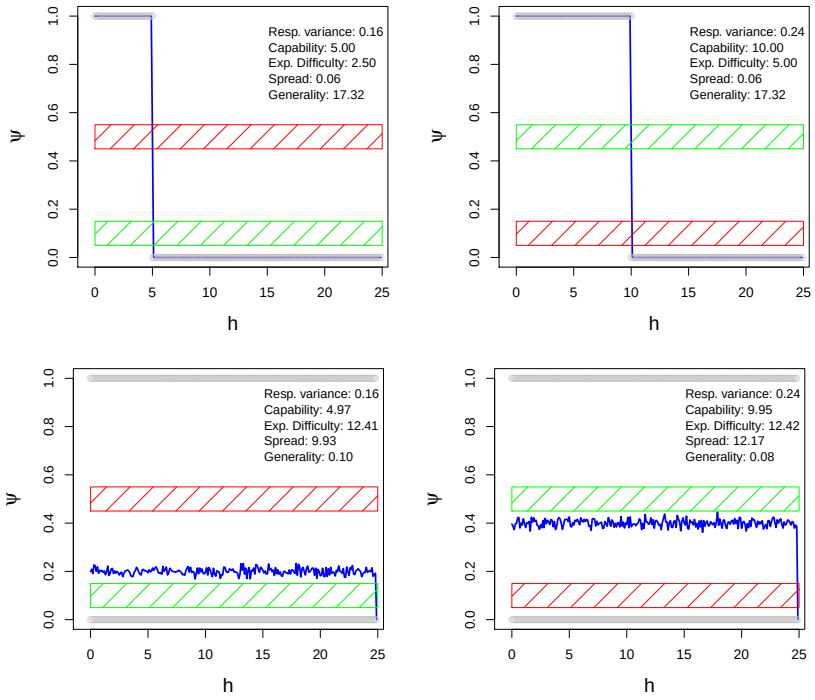

Figure 2: Four agent characteristic curves, with the bands showing two possible tests (red and green) grouping subsets of tasks. Top: We see two maximally general agents. Independently of how the groups are made for the two tests, provided the same range of difficulties is covered, the curves for each subset would be the same and so the effect on the populational generality. Bottom: groups can be made in such a way that the red test gets all positives for the bottom left plot but all negatives for the bottom right plot, and the opposite for the green test. As a result, the manifold might even be negative.

for group A was 0.47 but the mean correlation for group B was 0.78. Note that this does not mean that group A had worse results (in fact, it was precisely the group with highest average results), but rather that the *proportion of the variance* explained by g for the low-ability group was much higher than for the normal-ability group. This result was striking, especially if g is understood as general intelligence. It looked as if the more intelligent a population is, the less important g would be, in relative terms, to explain its variability. This observation turned to be known as Spearman's Law of Diminishing Returns (SLODR). The finding was replicated many times since then with different experimental settings [16, 14, 116].

Spearman looked for an explanation and found it in the *law of diminishing*

returns in economics. Many processes that are affected by many factors do not grow continuously as the result of the increase of one factor, so the influence of a single, albeit dominant, factor can become less relevant at a given point, being saturated. Spearman expressed it in this way: "the more 'energy' a person has available already, the less advantage accrues to his ability from further increments of it" [111, p. 219].

But this simile was not an explanation. Spearman postulated the "ability level differentiation", which considered that challenging items (those that can only be solved by the most able individuals) require the combination of many skills, and the prevalence of g would be smaller. Basically, for the easy items, the general intelligence or some general resources would be the only available skills for low-ability subpopulations. Detterman and Daniel [16] argued similarly that if "central processes are deficient, they limit the efficiency of all other processes in the system. So all processes in subjects with deficits tend to operate at the same uniform level. However, subjects without deficits show much more variability across processes because they do not have deficits in important central processes". Other explanations were introduced, such as that the "genetic contribution is higher at low-ability levels" [14].

Not only have the above explanations been put into question but the experimental evidence itself has been contested. One common counter-explanation of the phenomenon argues that it is not that g is less important for able subjects, but that they find many of the problems in the tests less challenging than the normal population so they are not forced to use general intelligence. They can solve the problems without (deep) thinking, i.e., more mechanically. In other words, the use of the same tests for both groups would be creating the effect. Relatedly, Jensen [73, p. 587] argued that the subgroups with higher abilities had lower variance than the subgroups with lower abilities. In fact, Fogarty and Stankov [27] performed an experiment where the more able group had to solve problems of higher difficulty whereas the less able group had to solve problems of lower difficulty. Under these conditions not only did SLODR vanish but even the more able group showed higher correlations.

This observation is more consistent with individuals having generality, such that if the distribution of difficulties of items is not adjusted for the two subpopulations, the items would be on the left of the steplike ACCs for many individuals of the more able group, so the correlations of the most able group would be smaller. Note that this would not appear for flat ACCs (with very low individual generality).

In other words, the SLODR, without adjusting the difficulties, would not appear if the individuals were not general at all. However, it appears if the individuals are highly general. And it is also easy to see that if we adjust the difficulties, so that the distributions are the same for both groups (and the relative distributions of abilities are the same), then we would have exactly the same manifold, so no diminishing or increasing returns.

Of course there is a pressure about resources when trying to achieve capability, and this may make the ACCs more compact for higher capability, having more individual generality for the more able group. That would entail an aug-

menting return, as postulated with the so-called Universal Law of Augmenting Returns (ULOAR) [40]. We will return to these issues under an evolutionary framework (pressure of resources) and also a computational framework, by looking at the invariance theorem and the stability of difficulty.

4.4 Individual generality, collective intelligence and the c factor

Finally, let us comment very briefly about collective intelligence [127]. It seems that maximum generality for individuals is not optimal for a group, as one individual will dominate the rest (agent transitivity), and the result of the group will be the result of the best agent in the group. With more specific agents, there could be more possibilities to go beyond the most capable individual. Of course, this depends on many assumptions about the dynamics of the groups, with the exact outcomes easier to derive when groups just combine their capabilities by voting or weighted voting (if confidence is used) [78, 77, 5]. We will explore this in the context of ensemble methods later on.

Also, the aggregation of several curves could be understood as a normal distribution on the reliability of the capability, transformed into a sigmoidal cumulative density function for the ACC. Consequently, findings such as the c factor [129], could be re-analysed by looking at the individual generalities first, rather than looking at the individual g scores (or IQ scores).

5 Evolutionary interpretation: generality and general intelligence in the animal kingdom

The study of intelligence in animals (including humans) usually distinguishes between domain-general and domain-specific kinds of cognition. Much debate has been held on whether or how much of these are present —or what the ranges are— in humans and other non-human animals, and how this relates to a modular view of the mind [26] or to a developmental domain-general learning [100]. It is also common to analyse whether social species are associated with more domain-general cognition, and the so-called social hypothesis (see, e.g., [74, 11, 122, 62]).

The definitions of what is general and what is specific also vary in the literature, but it is usually understood as coping with a wide range of cognitive tasks, or flexibility for changing cognitive demands in an unpredictable environment [103]. Note that this view is similar to the notion of generality we are discussing in this report, except for the explicit use of difficulty. In our case, we say that an animal or a species is cognitive general if it is able to perform equally well on a wide range of problems up to a limited difficulty. This is contrast to specific animals or species that display a smaller hardwired repertoire of domain-specific functionalities where they excel, but are unable to cope with tasks beyond the repertoire.

5.1 The manifold, the g and G factors and intelligence convergence in animal cognition

A data-driven approach to the issue of general intelligence in animal cognition has usually been conducted with populational analyses performed on several non-human species. Burkart et al. [10] provides the most comprehensive review to date of the study of the correlation manifold in non-human animals, both intra-species (denoted by g) and inter-species (denoted by G). The main conclusion is that "there is increasing evidence for g in nonhuman animals, particularly in mice and primates [...] At the interspecific level [...], studies of primates and birds provide a robust pattern consistent with G" [10]. Basically, if we represent the performance of several individuals or species for several domains, as shown in Figure 3, the evidence would be more in alignment with plot b, which shows that when one individual is good in one domain is usually good at the other domains, much in alignment to the early notions of general intelligence in humans and the positive manifold.

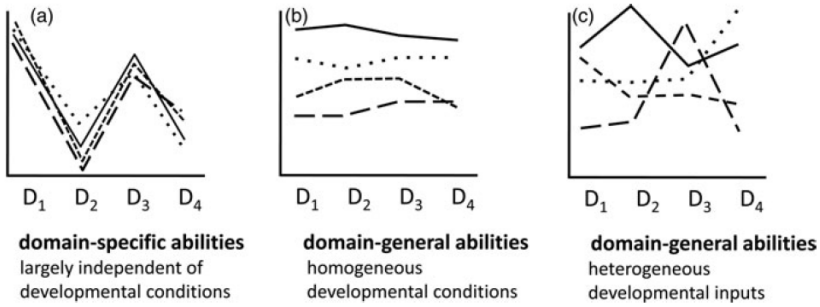

Figure 3: Three different possibilities for four individuals of a single species for four different domains D_1, D_2, D_3 and D_4. (a) The individuals behave better for some domains than others with very small differences between the individuals for each particular domain. (b) The individuals behave equally well for all domains, but some individuals show higher performance than others, also in a consistent way. (c) At the species level, there seems to be no difference between domains, but individuals perform differently for some domains, either by individual differences or by "heterogeneous developmental conditions". Copied from Fig.1 in [10].

So we are in a very similar situation to the human case. We cannot directly derive individual generality from these findings unless we postulate further assumptions, especially in terms of the difficulties used for the items in the domains. Of course, batteries are chosen such that there is variability of results to explain, so items of different difficulty are included. This variability is basically what is been looked after (a factor that explains a great proportion of the variance). However, it is not customary to perform a systematic analysis of difficulty (for instance, using cognitive demands for each item or using IRT).

Also, in the first place, the identification of domains (such that they are actually diverse) is one of the fundamental methodological issues in the analysis of general intelligence in animals. "The issue of task selection is thus closely linked to the identification of domains in animal cognition, which in fact is part of the empirical question that needs to be addressed in intelligence research in animals in general, by using batteries as diverse as possible and statistical procedures that are a priori agnostic to the underlying factor structure" [10].

5.2 Cognitive resources and generality

Still, the references to resources (cognitive demands) required for the tasks in several domains are usually part of the discussions. Burkart et al. [10], for instance, set the question around how much extra neural tissue is needed, taking into account that domain-specific cognitive adaptations may require much less additional expensive brain tissue [119] than domain-general cognitive ability, which is also less directly linked to fitness-relevant benefits. They face "the puzzle that domain-general cognitive ability apparently evolved in at least some lineages, or perhaps even in birds and mammals in general, even though its evolution has had to overcome more obstacles compared to the emergence of domain-specific cognitive adaptations" [10]. One possible theory that explains this puzzle is the *cognitive buffer hypothesis* [2], which states that this extra effort in domain-general cognitive processes in larger brains buffers animals against environmental variation, and pays off for a wider range of behavioural patterns given by innovation, learning and, most especially, cultural transmission [58, 119, 80, 107].

Evolution usually finds a trade-off between specialised functions and more general capabilities, according to the effort that has to be put in terms of evolutionary innovations and energy consumption of bigger brains on one hand and how expectable and regular the tasks that are faced by the species are in their environments. We can see this trade-off in Figure 4, where we compare the gains and the efforts of a domain-general cognitive enhancement versus a domain-specific cognitive enhancement.

Of course, how meaningful the specific numbers are depend on how well we can estimate the effort for general solutions versus specific solutions and how likely the specific tasks are versus all other tasks. Actually, Figure 4 assumes that all tasks are equally likely (or, more precisely, all difficulties are equally likely). When the probability of some specific behaviours or domains is very likely in the environment of the species, then specialisation will of course pay off. It is when there is environmental unpredictability that many tasks are similarly likely, and then the pressure for more general intelligence takes us to the kind of increase like the violet band in the figure rather than the orange one. Actually, in an environment where most tasks change in a few generations we would have an ACC closer to the maximally compacted one and maximum generality (as given by the compactness property, #2), as this would be resource-optimal in order to obtain maximum capability (and maximum success if tasks are so unpredictable). Of course, many tasks requiring cognition, such as navigating

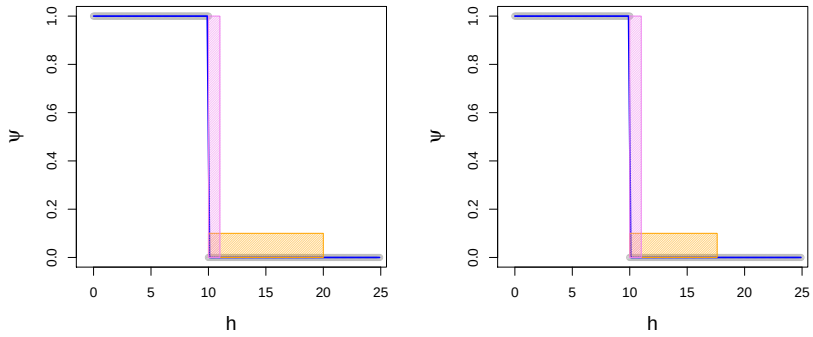

Figure 4: Using ACCs to represent two different ways in which the capability of a species can be enhanced, with a domain-general cognitive enhancement (vertical violet rectangle) or a domain-specific cognitive enhancement (horizontal orange rectangle). Left: both rectangles cover the same area, and hence increase capability in a similar amount. Right: both rectangles imply the same extra effort ($m_j = \int_{10}^{11} h \cdot 1 \, dh = 10.5$ vs $m_j = \int_{10}^{17.6} h \cdot 0.1 \, dh \approx 10.5$, according to Eq. 2), but the specific one (orange rectangle) now has a smaller area, and hence less increase in capability than the general one (violet rectangle).

and eating —but not foraging or hunting— might still be linked to a few particular specific skills, as they are more constant in the evolutionary history of many species.

How difficult a domain-specific functionality is or how much effort it requires can be analysed in different ways. One first way is to look at the energy effort, by examining the involved neurological modules that are dedicated for that functionality, and map this with energy consumption. A second approach is to estimate evolutionary effort by looking at changes in DNA that make the functionality possible (from an ancestor that did not have it), contrasting with the ecological pressures and other similar functionalities. A third pathway is to set these tasks in an isolated or abstract way and make them be learnt by systems that do have general capabilities, and estimate their difficulty from them. In this case, extreme care has to be made for many confounding factors. Finally, a fourth possibility is to determine the difficulty of tasks intrinsically (e.g., working memory requirements, pattern complexity, etc.).

The analysis is complex, but in many ways it is what research in animal cognition has focused on in the past decades. Without getting into these estimations, many studies about general intelligence would just conclude that many animals have a balance of general-domain and specific-domain solutions, which, in general terms, is not very surprising. The challenge is to determine which or at least how many cognitive skills are specific or general, or going beyond the dichotomy of domain-specific versus domain-general [72], how much general a

species is according to the range of tasks it solves and their difficulty. ACCs can help analyse this visually while a notion of generality can help us analyse this numerically, and separate this from capability. At the moment, [10] focuses on whether the species has g and how its results may compare with other species (through G and a comparison of magnitudes). But most of the discussion in [10] and its responses turns around the question of whether the correlations might be a produce of something that is not general intelligence. This has its roots in an inadequate definition (or multiple different definitions) of general intelligence, which in some cases is linked to results on tasks and in other cases it is linked to processes. Indeed, the concept general intelligence conflates magnitude and distribution, as the definition of 'intelligence', either explicitly or implicitly, integrates a minimal degree of generality.

5.3 Looking at evolutionary selective pressure through observable scores: capability and generality

An alternative way of looking at this is in terms of two observable indexes: capability and generality, especially if we see that some less general species are able to solve very complex problems by specialisation that other more general species cannot do. Plotting generality and capability against the level of social interaction (intra-specific and by diversity of predators), cultural inheritance, neural tissue mass, etc., with octopuses, hienas, koalas, racoons, primates and corvids, among other species, is expected to scatter points on very different locations. As a result, this could also help us see whether these traits are related, or whether there might be one-directional causalities. Both capability and generality are observable variables, the first is aggregated performance (the area under the ACC curve, eq. 1) and the second is a metric of how compact this performance is (how steplike it looks over difficulty, eq. 5).

Figure 5 shows a simulation where 200 individuals are generated on random results on 200 items, and we see different selective pressures on the capability and the cognitive effort on the individuals. As we see, not only do we get more generality but also the correlation between capability and generality increases, due to the pressure on minimising effort (while keeping or maximising capability).

Figure 6 shows another similar simulation where items have a uniform range of difficulty and success for the tasks is randomly proportional to the difficulty. In both figures the correlations can get very high since the pressure goes in the same direction: more capability and less effort. This is simply the result of the compactness property (#3). But again, it is important to notice that as generality and capability become more correlated (especially in humans) there is a tendency in confounding them, and start talking about *general intelligence*, without knowing clearly whether the emphasis is on generality or capability.

The theory of general intelligence, the positive manifold and the g and G factors have all (in different degrees) raised bitter controversies. Setting aside the interpretation issues, one of the major arguments against these theories is that they might be considered statistical artefacts, produced as the result of

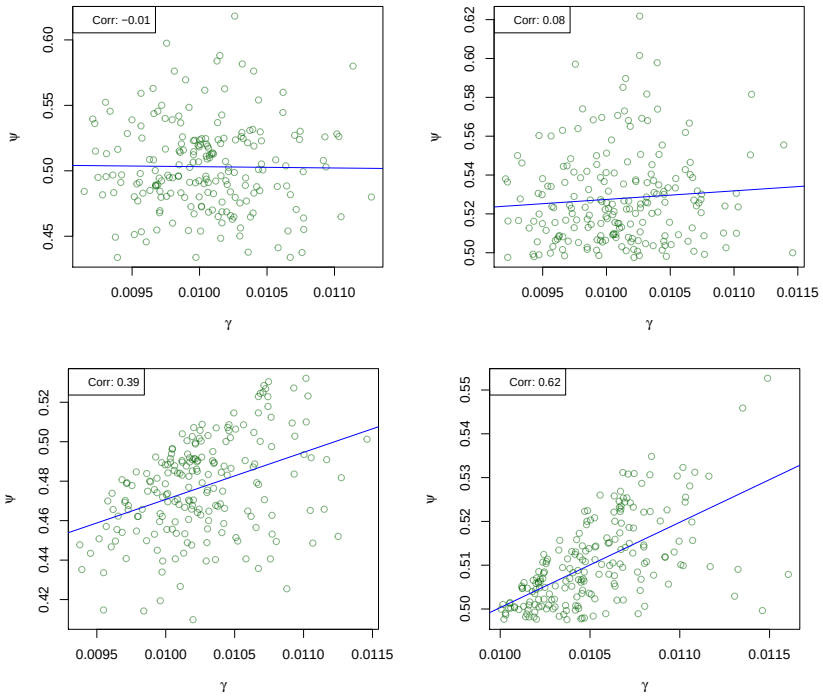

Figure 5: Distribution of capabilities and generalities of a simulation where 200 individuals are generated on random results on 200 items. The top left plot shows the original case with no selective pressure. This is not very interesting as all the ACCs are flat. The top right plot performs a selection per capability, where only those individuals with capability greater than or equal to 50% survive. The bottom left shows a selection by effort, where only those individuals that require less than 100% over the minimum possible effort (a maximally compacted ACC) survive. Finally the bottom right combines both selections at the same time.

making some choices on the items and test batteries, such that they fit the population of individuals (not too easy, not too difficult, so there is variance to explain). In a very insightful way, [128] break the species groupings by considering humans and chimpanzees together into a single population and then correct for these "ceiling or floor effects", by reducing the number of tests to those that have higher coefficients of variance. Figure 7 shows the correlation of scores (d) and g loadings on the y-axis against different values of the variance produced by progressively selecting the tests with higher variance. Although not mentioned in [128], this analysis is of course closely related to the SLODR (and the alternative ULOAR hypothesis) discussed in the previous section, where by

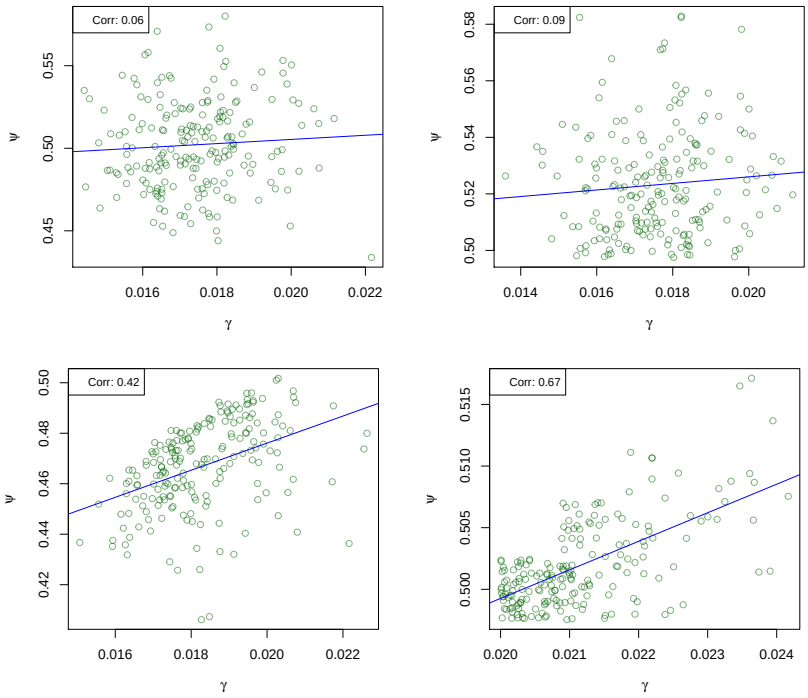

Figure 6: Same as Figure 5 but with examples of different difficulties (ACCs are triangular originally). Top left: no selection. Top right: selection by capability. Bottom Left: selection by effort. Bottom Right: selection by effort and capability. For the two bottom plots, maximum effort set to 25% over the minimum possible effort (a maximally compacted ACC).

adjusting the variance we can get that g and scores can grow together, as we see in Figure 7.

Of course the criticism about how tasks and individuals are chosen or split will always be around, as these constructs are populational (on the tasks and the individuals), and there seems to be a chicken-and-egg problem. By looking at generality, as an individual observable measure, we can simplify the analysis in many ways: the measure does not depend on a population of individuals (no need to arrange them into species or groups) and it is algebraically independent from capability (of course unless an evolutionary or other kind of efficiency pressure is applied). For instance, actual plots like Figures 5 and 6 can be used as an alternative to Figure 7, and done for individuals of many species together.

Another kind of criticism around the study of general intelligence is about whether "the positive manifold provides little or no constraint on the possible architectures of cognition" [104]. General intelligence may then originate from

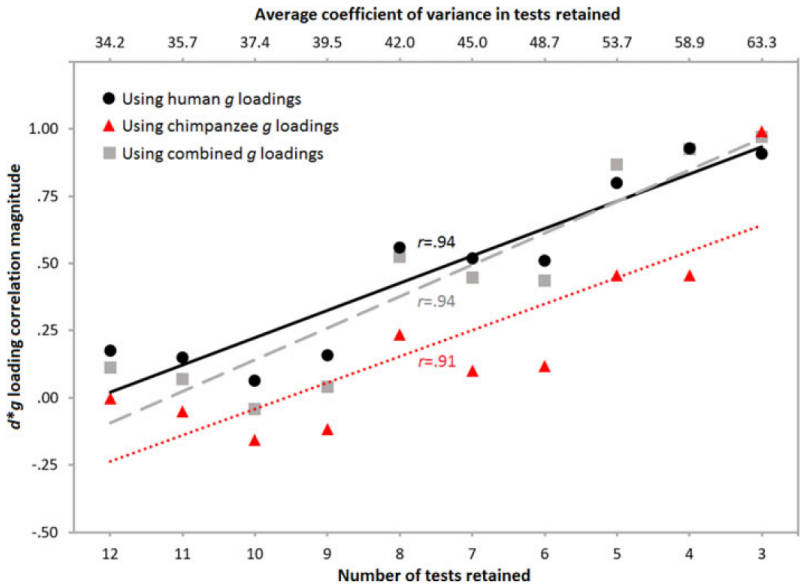

Figure 7: Correlations between task g loadings and the scores d on the y-axis as a function of the average coefficient of variance in the tests retained, choosing them by removing those with smallest variance first. Trends shown for chimpanzees, humans and a combined population. Copied from [128].

primary specific modules being boosted by more general secondary modules (or evolved in this more compressed/abstract way for the economy of the brain), by a wide range of specific modules that are switched on or off depending on the task at hand or by a truly general system helped by particular biases according to what environment demands are most frequent for a species. All this diversity of explanations could be extended to generality, as a high value of γ can be obtained in many ways (but not as many as g, as we discussed in the previous section). As we will see in the following section, looking at individuals that have gone through no selective pressure, or a different one (i.e., AI systems) can give us a wider theoretical and empirical scenario to exclude some interpretations of the existing findings and especially on new research looking at the values of γ and ψ in the animal kingdom.

6 Computational interpretation: generality and artificial (general) intelligence

The debate along the spectrum between general intelligence and specific (or narrow) intelligence has also pervaded artificial intelligence since its inception. The very early attempts were directed towards a General Problem Solver [96] and the goal of "generality" [24]. In the following decades, many of these programs

failed to fully realise the complexity of intelligence, while other more narrow applications started to be successful.

In 1978, John McCarthy published a new version of his 1971 Turing Award Lecture on "Generality in Artificial Intelligence" [91], recognising that one of the major problems was that if behaviour was represented by programs, then these programs could only cover a finite set of domains or problems.

Of course, these were the times were machine learning was not a dominant paradigm in artificial intelligence. Nowadays, the use of machine learning techniques, coupled with sufficient data, allows systems to be adapted to different domains, using the same algorithm, which *generalises* the data. Generalisation is an intrinsic —if not definitional— part of learning. Learning is hence the way in which AI systems (and human and non-human animals) can adapt to unseen situations. In other words, when considering a large and diverse number of tasks, coding particular solutions for all of them is infeasible, and hence learning becomes the solution.

Consequently, it may seem that (machine) learning systems are then general by definition: give a learning system sufficient examples and it will learn any possible task. The goal of machine learning, and AI, would be to define this universal machine learning system. While this idea is still behind some of the narratives in machine learning and artificial intelligence, there is an important objection to this universal generality: *efficiency*. Some systems can potentially learn any function, given a sufficiently large number of examples. The question is how many examples, how much time and how large the model might be. The *answer* to –or *cause* of— tthis problem is known as *bias*. By embedding a particular bias for a learning algorithm, one can accelerate learning for some problems while making it harder for some other problems.

There are many ways of explicitly or implicitly introducing strong bias to a learning algorithm: specialised architectures, hyper-parameters, background knowledge, and the very algorithm itself. By using these particular biases, we can have AI systems that can solve particular pockets of problems: speech recognition, machine translation, robot navigation, medical diagnosis, face recognition, etc. Interestingly, by a shrewd use of more and more computing power, some of these algorithms are requiring less physical time (and occasionally fewer examples) to learn these tasks, approaching, at least in some areas, the flexiblity of some animals.

Still, there is a view that artificial intelligence does not produce general systems. Even if the same deep reinforcement learning can learn to play Go or Chess by just changing the rules [106], the *same* algorithm cannot learn to navigate a room. Of course, there are algorithms that can learn to navigate a room and have similar principles (and even shared modules underneath). However, they need a great amount of hyperparameter tuning, input and output transformation, and other changes to the architectures and the optimisation operators to make them work for a different domain.

Because of all this, the area known as Artificial General Intelligence [1], where the *same* system must be able to solve a range of problems, is still seen as a counterpoint to a bevy of systems that are successful for more narrow

domains, even if they are fuelled by machine learning, and built upon general principles looking for abstract representations.

Unfortunately, to the dismay of some members of the AGI community, the term AGI is now commonly used as synonym of ill-defined buzzwords such as human-level machine intelligence, human-level artificial intelligence or even superintelligence, without a proper analysis of what the 'G' in AGI actually means, and how it can be distinguished from mainstream AI [8].

6.1 Generality and all possible tasks

The reduction of AGI to anthropocentric views of intelligence has an intuitive appeal. We are interested in those tasks humans can solve. But which are these tasks? Or, more conspicuously, what are the tasks that humans —the hominids characterised by their general intelligence— cannot do? We can analyse this question and put the notion of generality to its limits by considering *all possible tasks*. One possible way of doing this is by defining the set of all computable tasks, where tasks can be framed in a testing scenario, where agents can learn from experience. In other words, we can consider all possible learning tasks (see [108, 109, 124, 33, 63, 41] for different ways of doing this). Apart from the particular formulation and setting, the most relevant feature is how to distribute a weight or distribution over all possible tasks.

Let us start with Solomonoff, who defined all possible sequential prediction tasks and an associated distribution, the algorithmic probability [108, 109]. The set of tasks is just defined by the problem of estimating the next bits of all the sequences that can be produced by a universal Turing machine UTM. While all sequences are generated, their distribution (the algorithmic probability) depends on the reference UTM. In a way, this was an elegant way of representing the notion of bias in machine learning. Depending on the chosen UTM, some concepts will be easier to learn than others. Still, the great contribution by Solomonoff was that he showed that the same algorithm can be used for all UTMs (biases), and convergence can be obtained. A universal learning algorithm exists, it always works, but it will work more or less efficiently depending on the chosen bias, the reference UTM. In other words, each UTM assumes a prior about the world, and observations whose underlying pattern is simpler for the chosen UTM (smaller Kolmogorov complexity) are more likely than those observations with more complex patterns. Solomonoff integrates Occam's razor and Epicurus, as his theory considers the combination of all theories that are compatible with the evidence, weighted by their Kolmogorov complexity.

On the other extreme for the choice of a distribution we find the assumption that every possible problem's output is equally likely. In a sequential prediction problem this would be expecting all sequences to be equally likely or, in classification problems, to consider all combinations of inputs and outputs equally likely. This is technically known as "block uniformity" [66], with the uniform distribution being a special case. Under this assumption, we have the conditions for the so-called no-free-lunch theorems [126, 124, 125], leading to the conclusion that, on average, no method can be better than any other. A general-purpose

learning system and hence the very notion of 'general intelligence' would be simply impossible [21]. Moreover, every agent would solve exactly the same number of tasks, so there would not be any variability in capability, effort and of course generality.

The NFL theorems are very relevant, because our observation that learning systems exist and work (in animals and computers) can only happen if the assumption is not true. This is a pragmatic or ab absurdum rationale, but there are more epistemological ones: choosing all perceptions as equally likely is difficult to reconcile with a world with physical laws and other agents around (plants, animals, conspecifics) that do not behave randomly. Actually, if we consider all these subsystems computable, Solomonoff's view is more natural, as the output of a UTM fed with random bits is not random. In other words, what we perceive, our world, is filtered through many machines, making those patterns that are produced from systems with limited resources more likely.

From this view of all possible tasks, one can define a (universal) distribution according to the complexity of the generator of tasks. However, one can also define the distribution by looking at the complexity of the solution for the task, which can be seen as its *difficulty*. This way of representing/weighting solutions by their difficulty is common in psychometrics, but was first introduced in the context of all (sequential) tasks in [33]. When one goes from sequential tasks to interactive tasks (such as reinforcement learning [35, 71]), the difference between the smallest program that generates a task and the smallest program that solves the task becomes conspicuous. Setting the distribution according to the former led to the notion of *universal intelligence* [81]. Setting the distribution according to the latter led to the notion of *policy-general intelligence*, assuming a uniform distribution over *solutions* for each task difficulty [38, 37, 41]. We can see some of these choices in Figure 8.

6.2 The choice of diversity and difficulty

The important thing about a theoretical account of all possible tasks, and especially if we know how we generate them, is that we can control for two things that are crucial for generality: the diversity and the difficulty of the tasks. If we look at diversity first, the schema on the top of Figure 8 makes it very difficult to ensure that the set of tasks is going to be diverse, as we generate tasks according to a distribution on their definition, but not about their solutions. Apart, if the choice is a universal distribution as in [81], then the distribution is dominated by a few tasks coping most of its mass [59, 42]. For the schema in the middle of Figure 8, we have at least some range of difficulties but, still, that does not ensure that the solutions might not all end up being of the same kind. Finally, it is the choice at the bottom of Figure 8 that ensures diversity by the most entropic choice of a distribution per each difficulty (assuming the number of solutions per difficulty is finite). This choice is the uniform distribution.

For instance, Figure 9 shows an ACC where instances have been generated according to the bottom schema in Figure 8. If we consider all difficulties as equally likely, and assume the curve is 1 for $h < 7$ and 0 for $h > 14$, then we have

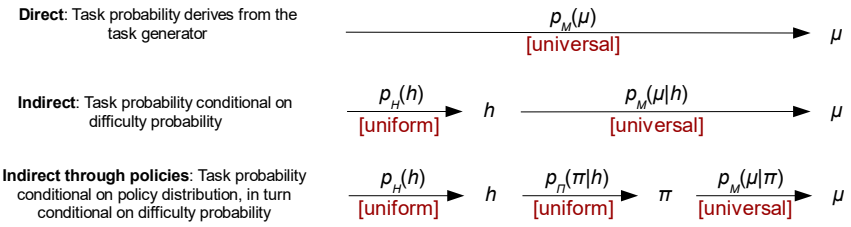

Figure 8: Different ways of generating tasks (or deriving their distribution). Top: the probability of a task is given by its generator. If the number of tasks is infinite, but countable, a uniform distribution is not a viable option, and a universal distribution must be used instead, making this equal to *universal intelligence* [81]. Middle: we first define a distribution of difficulties and then we define tasks according to that difficulty. In cases where the difficulty of a task can be derived from the definition of the task, this is a good option, as in [56, 33]. Bottom: again, we first define a distribution of difficulties and then we derive solutions matching that difficulty. Finally, tasks are generated according to the solution. This is actually an option when the definition of a tasks does not say much about the difficulty of the solution, such as interactive tasks, as used in [38, 37, 41]. Note that for the two last ways, if the difficulty distribution is uniform, the expected success on a random task drawn from the distribution is equal to the area under the ACC, which is capability, as for eq. 1. (Image adapted from [41, Fig.9.7].)

the ACC shown in the figure, with capability $\psi = 9.86$ and generality=0.39.

A theoretical view also allows us to consider different options for difficulty. Having all tasks sliced by difficulty provides us with a way to understand the success of an individual in relation to the resources used. For instance, if we consider *difficulty as the complexity of the simplest solution*, there are few interesting consequences. First, we have that for every agent, there is a difficulty for which its ACC is zero, so the area is always finite. Second, we can precisely determine how many solutions of a given difficulty there might be. For instance, we can calculate the resources according to different situations:

- We can consider difficulty as the length of the solution with lowest Kolmogorov complexity, i.e., $\hbar(\mu) \triangleq \min_{\pi:A_\mu^\pi=1} L(\pi)$ where $L(\pi)$ is the length of the solution π. Then the number of solutions for a given difficulty h would be 2^h. In this situation, we can derive from the compactness property (#2) that the optimal curve is again one with $\gamma = \infty$. To achieve capability ψ, a non-learning system having predefined solutions for a large number of tasks would require a minimum of $\sum_{h=0}^{\psi} h2^h = (\psi-1)2^{\psi+1}+2$ bits, plus the necessary code or neural wiring for making the switch among the $2^{\psi+1}-1$ solutions (assuming the solutions have nothing in common, because exhaustiveness here makes it difficult to compress this into a more hierarchical or reusable architecture). According to this situation, we can

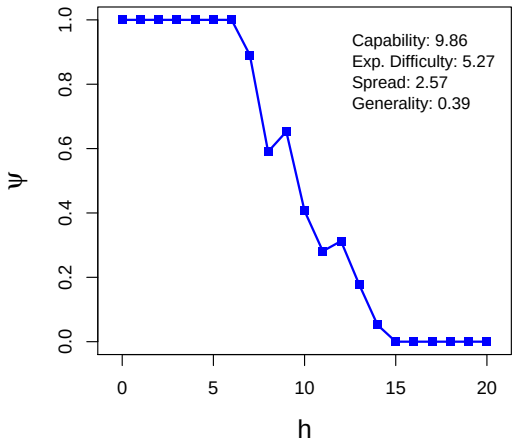

Figure 9: Average human results on exercises of different difficulty (h) in the C-test [56, 33], with the derive metrics shown on the plot.

see that the "size" of the "brain" would grow more than exponentially. On the contrary, if instead of predefined solutions, we consider a learning system, the size would be reduced as much as we would like, but we would need to consider the availability of data and the learning effort instead.

- We can understand difficulty with Levin's Kt complexity, as advocated for in [56, 33, 39, 50, 41] because of its connection with Levin's optimal universal search [82, 83]. In this case, we define $LS(\pi, \mu) \triangleq L(\pi) + \log S(\pi, \mu)$ where L is the length of the solution π and S the computational steps μ uses to solve the task π. Difficulty would be $\hbar(\mu) \triangleq \min_{\pi: A_\mu^\pi = 1} LS(\pi, \mu)$. With this, we could still consider that the number of solutions for a given difficulty would be less than (but still approximately or linear with) 2^h. The result for a non-learning system would be then similar, but now we will have to take into account the time to determine which problem we are facing, which must choose between $2^{\psi+1} - 1$ solutions. On the contrary, the result for a learning system using this schema would just simply be the expected difficulty $\mathbb{H} = \psi/2$. This is measured in the logarithm of computational steps[9], so the expected computational steps using a universal search would be $2^{\psi/2}$.

From the above, we see the difference between a system with a predefined repertoire of solutions and a system that learns those solutions[10]. Even if the above

[9] These would be the *bints*, as used in property 7 in section 3.3.

[10] Note that this is not related (and also looks apparently opposite) to the distinction between learning tasks and knowing tasks in [118].

ignores the training examples or interaction needed to learn the concepts, we see that there might be a trade-off between pre-wired and learned solutions, depending on the size limitations and the speed of the system.

The examples above are important to clarify the distinction between nature-vs-nurture and general-vs-specific. Whereas we have the tendency to associate inherited functions with specific functions, this does not have to be the case a priori, according to the definition of generality we are considering here. This may be a consequence depending on what resources are most relevant. Note that in the two analyses above, we derive the minimum resources following the compactness property (#2). Assuming all difficulties equally likely, one should focus on those policies that require fewer resources. Of course, if some particular pockets of problems of high difficulty are more likely than many problems of low difficulty then there is a rationale to cover those pockets specifically, so having less generality.

In all these cases we are using a distribution of tasks that is not based on a particular species or environment —they are not the tasks a human or animal would find in their lifetime. Accordingly, these distributions can be criticised as arbitrary. However, it is not true that all humans (and much less all animals) face the same fixed set of tasks. Precisely because of this, many psychometric tests include very abstract tasks, in an effort to be independent of particular human groups, and some (like Raven's matrices) may even look very unrelated to the natural (ancient or modern) environments humans face. However, it is well known that IQ tests lack measurement invariance when applied to other groups (e.g., people with some disabilities, children, etc.), non-human animals and, most especially, computers. In the latter case, it is not that they are particularly unfair for computers, but that AI systems can specialise for these tasks [18, 7, 55, 88]. In a way, we can get generality inside the test, but a high specialisation to the tasks that are outside the test. Restricting to a particular kind of tasks facilitates systems that specialise on them, and this is particularly exploited in AI.

Hence the idea of using all tasks. Still, how much will the task distribution depend on the representational language or mechanisms used to derive the set of tasks? The invariance theorem, independently introduced by Solomonoff, Kolmogorov and Chaitin (see, e.g., [84]) says that any universal representational mechanism (language) can code any program as efficiently (in size) as any other up to a constant that is bounded by (but generally smaller than) the sizes of the definitions of both languages. This makes the concept of Kolmogorov complexity machine-independent, at least to an additive constant factor.

However, the definition of "universal intelligence" [81] has been criticised by this dependence on the reference machine, which is actually leading to different definitions according to what UTM is used to generate the universal distribution [59, 42, 41]. The main reason is that the invariance theorem appears in the exponent of the distribution ($2^{-K(x)}$), and the additive constant becomes an exponential one. In contrast, the two versions on the bottom of Figure 8 put back the invariance theorem as an additive constant on the scale of difficulty. This means that the scale upon which all other measures are derived is relatively

more stable. For instance, given the spread for an individual using a notion of difficulty on a reference machine, then this spread will be at most increased by a constant that does not depend on the individual. Also, as capability grows, the invariance theorem starts having more relevance. This can also be seen as the issue that systems with very limited resources (or capabilities) will be more dependent on the reference machine.

Still, using two different reference machines might lead to very different x-axes for the ACC and hence different capability and generality scores, which is of course what underlies many discussions about whether tests are biased against or in favour of a group. But there are many "bias equalisers", especially in testing, that can be used to determine capabilities and generality more independently [45, 44, 20]: 1) introducing a testing apparatus that is novel for all subjects, 2) analysing groups after they are raised in or adapted to the same culture or using the same language, 3) present problems that have to be solved by combining or using a set of constructs or elements that are abstract and new. These procedures are common in animal cognition and human intelligence testing, but not that much in AI research [36, 69, 67].

In practice, we do not need to consider all possible tasks to derive metrics of generality in AI. We can do this for any test battery or benchmark for which we are interested in deriving the generality of a particular AI algorithm or agent, be it in machine learning, planning or machine translation. In order to start we only need a metric of difficulty. It does not have to be a universal metric, as described above, but a customised one instead. It can be derived in many ways:

- Anthropocentric difficulty: we can use human performance as a reference for the difficulty of a set of tasks. This can be obtained as an indicator that is inversely related to the success of average humans in each task.

- Populational difficulty: this can be derived by using a population of AI techniques for the range of problems. For instance, [90] apply IRT to derive the difficulty of machine learning instances. This idea can be applied to datasets and other kinds of problems in AI (e.g., the ALE benchmarks, [89]).

- Intrinsic difficulty: any meaningful characterisation of difficulty can be used here. For instance, the difficulty of a planning problem can be based on a series of features about the problem:, such as the number of components, its structure, the degree of noise, etc. Note that difficulty is different from computational complexity, but time complexity may be an important factor.

- Integral difficulty: some other notions of difficulty can integrate space resources, computational time, energy consumption, data required, etc., especially when including very different tasks. For instance, [101] aligns difficulty with the number of trees used by a random forest classifier, providing a very clean mapping to resources and effort..

- Opponent difficulty: in those cases where other agents compete or cooperate, we can use the capability of the opponents (or a measure inversely related to the capability of cooperators). Note that this makes this option populational as well.

In general, whenever an evaluation procedure is established in AI, there is a selection of tasks from a certain domain and for a particular range of difficulties. For instance, one rarely finds Hofstadter's "Gödel, Escher, Bach" [61] as an instance for a machine translation benchmark. It is too hard to be discriminative for AI. Usually, the benchmark tasks are selected to cover an application area (usually of scientific or industrial interest) and the difficulty of the items is chosen such that they are neither too easy nor too difficult for the state-of-the-art algorithms. This is natural, but this is implicitly assuming a type of ACC nobody checks in the first place, and a very malleable notion of difficulty, adapted to the situation. This also makes the analysis of progress in AI hard to assess, as the tasks in the domain and their difficulty are changing, like a moving target.

6.3 Generality in competitions and benchmarks in AI

These options for difficulty can be applied to an increasing range of AI competitions and benchmarks [51], especially those that are aiming at more general-purpose AI. Some of these are the general game playing AAAI Competition [28, 29], the reinforcement learning competition [123, 17] (which featured the 'polyathlon', with several domains), the genetic programming benchmarks [92, 121], the general video game competition [102, 99], and the arcade learning environment (ALE) [6, 102], a collection of Atari 26000 video games, which "has incentivized the AI community to build more generally competent agents" [86]. It is important to note that the introduction of new platforms and benchmarks where hundreds of tasks can be potentially be implemented [12, 52] is not usually accompanied with a verification that the agents that have highest performance are also more general. Recognising that the diversity and difficulty of the tasks must be explicitly determined is one important outcome of our analysis so far, and a metric of generality in these terms would help to flesh out.

From all the ways in which difficulty can be introduced, and hence generality can be obtained, we are going to illustrate the last case of the bullet list in the previous subsection ("opponent difficulty"). This choice is motivated as it seems less evident than the rest and has interesting connections with competition ratings and some particular notions of transitivity. In particular, we are going to analyse the results of the World Computer Chess Championship (WCCC), usually part of the Computer Olympiad, where several computer chess players compete against each other. Figure 10 shows the ACCs of the participants of the 2005 and 2015 editions, taking the score of the opponent as difficulty (if two or more opponents ended up with the same score, they are considered together as "tasks" of the same difficulty=. This is why we have values on 1 (wins), 0.5 (draws), 0 (losses), but also some other values.

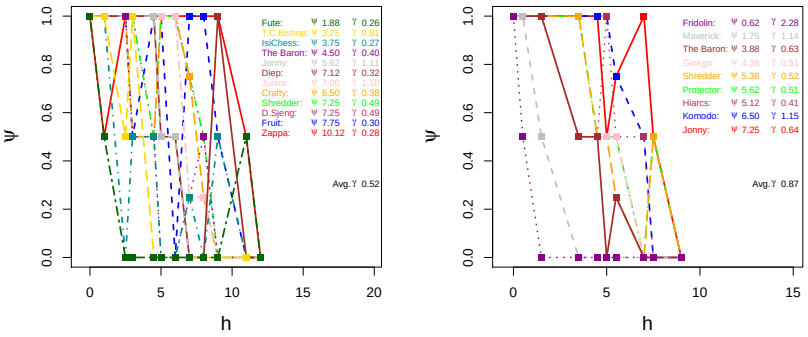

Figure 10: ACCs for all the participants in the World Computer Chess Championship using the final score of the opponent as difficulty. Left: Reykjavik 2005 with 12 participants. The winner (Zappa) and the last one (Fute) won and lost all matches respectively except the one between them, which was surprisingly a draw. Right: Reykjavik 2015 with 9 participants. Here, no low-rank beat a high-rank participant, and draws were usually between participants with close scores. Accordingly, the average generality is much higher in this case. Data from https://www.game-ai-forum.org/icga-tournaments/game.php?id=1.

We see in Figure 10 that curves are decreasing and generally quite steep, and generalities are relatively high. From the aggregate numbers we could conclude that the 2015 edition has more general players. Nevertheless, we have to be a little bit more careful, as the notion of difficulty here is populational and both populations are not the same (not even in number of participants), so they are not comparable. This is one of the issues of using populational notions of difficulty.

In this particular scenario, we can also conclude that the degree of transitivity is high, in light to the generality values. Note that in adversarial settings like chess, agent transitivity and task transitivity are two sides of the same coin, as tasks are opponents, which are also participants. Of course, there are more specific metrics to produce scores in tournaments (e.g., the Elo rating [22] or more sophisticated schemes [4]), but this example illustrates how to apply the individual generality score to a situation where the difficulty of a task depends on other agents taking place in a competing or cooperating role, which is especially necessary for social and adversarial situations [43, 68, 70].

This is also especially interesting for systems that improve with self-play, like AlphaZero [106]. In these settings, it is important to check that the system does not get better and better against more competitive opponents but may end up losing (or drawing more frequently) against weak opponents. This leads us to the more general question of whether a system that develops over time becomes more or less general [53, 87]. As the system evolves, we may experience less flexibility but a wider covering of tasks, and this can be studied using metrics

of capability and generality. It is also interesting to analyse those tasks where transitivity may be more problematic (and hence generality). For instance, matching pennies [60, 54, 47] is a game where a random agent has an expected result of 0.5 with whatever other opponent, resulting in a very flat ACCs. If many of these agents are included in a population of opponents, then it will become very difficult to attain generality.

In the end, the progress of some techniques in AI can be made in such a way that generality is preserved, and the ACCs are just *translated* to the right as the technology improves. We now have tools to check whether this is the case, or some new techniques solve more challenging problems at the cost of being worse at simpler problems. This is particularly relevant as progress in AI can be attained by combining several approaches, in areas such as ensemble learning or portfolios, where a big switch approach determines which technique is most appropriate for a particular instance. This modular approach to solving problems may well end up in specific solutions and creating gaps, where some relatively simple problems are not solved, with lack of generality. But this modular approach, combining many specific solutions, if the set of tasks remains constant, may increase generality (and capability), especially if the combination covers more of the easy ones than the difficult ones. Again, we see that generality measures how capability is distributed in terms of difficulty, but it does not impose constraints on how this is done. It may even include human computation, collective systems, cognitive services or hybrids, in the same way that humans can be enhanced by personal assistants or other devices and increase or decrease their generality because of this.

Of course, if a modular solution requires hundreds of specific subsolutions, the cost of keeping all them and designing an appropriate and efficient switch to determine which one to use may end up being less efficient than a more integrated solution, as we have discussed above. The relation between generality and resources is another way of looking at compression and generalisation, Occam's razor, the MML principle, etc., in machine learning, genetic programming and other areas in AI (e.g., [120, 79, 19]). Actually, the issues of generalisation and difficulty were usual (although from a different perspective) in the early days of genetic programming [76] (using the term 'generality' as 'generalisation power' or 'avoiding overfitting'). Commonly, the notion of generalisation is usually linked to whether a model extrapolates from the training data to the test data, and a proper validation will just equate this with performance. But, generality, as introduced in this report, just measures the distribution of success across difficulties, and can be applied to learning problems, planning problems, deductive problems, more in the original spirit of McCarthy [91]. This is in the end related to parsimony in scientific theories [48, 34, 46, 49] and even software systems [57].

7 Discussion

We started with the implausible assumption that generality can be seen as a cognitive system behaving well on a wide range of tasks, *independently of their difficulty*. While this might be the case for theoretical, idealistic, agents (e.g., AIXI [65, 64]), it fails to accommodate the fact that resource-bounded agents will necessarily fail on an infinite number of tasks, simply because they are beyond their capacity. From here, we could conclude that any system specialises for the subset of problems that are easy according to its resources. Comparing degrees of generality would then be a chimera.

We can escape this apparent contradiction by putting difficulty at a foremost place from the very start, trying to derive measures of difficulty that are independent, or at least sufficiently invariant or fair, to the agents that we want to evaluate. It is not surprising that the notion of difficulty has a prominent role in psychometrics, and it is also pervading the evaluation of non-human animals, either deriving from the analysis of the cognitive resources needed (working memory, size of the solution, etc.) or emanating from populational approaches (e.g., using IRT).

Once we establish a metric of difficulty for a range of tasks, and we see the results of an agent as an agent characteristic curve, we see that the notions of capability and generality appear as the two most descriptive indicators to summarise the curve. If the results follow a monotonic decreasing function, these would correspond to metrics of location and slope respectively, as observed from several models in item response theory.

However, we have seen that, because it is so unnatural to think that an agent can score equally well for easy and difficult tasks, most approaches in the analysis of human and non-human animal intelligence somehow assume this (decreasing) monotonicity in the process, and the whole analysis ends up mixing capability and generality. Actually, if an agent is shown to have poor generality, this is usually seen as a problem, a bad fit to the models, and something that should be corrected. In a way, there is some circularity if we try to analyse general intelligence (and derive the g factor) and at the same time one we assume that agents are going to show a (decreasing) monotonicity between difficulty and response.

By decoupling measures of generality and capability from the beginning, we can actually see that there is variability in generality that is to be explained as well, and can be done at the individual level. We can also analyse where the generality and capability values locate for a particular individual (be it human, non-human animal or machine) and then, and only then, study how it evolves collectively (as a group, population or species) or in terms of development. Once this is done, we can finally analyse that if resources are a (selective or designing) pressure, then the compactness property (#2) says that generality will appear in several situations, as we see in the animal kingdom, and more incipiently in AI.

As discussed elsewhere [44, 41], the evaluation in AI is now facing some of the challenges the evaluation of intelligence for humans and animals have faced for

over a century. However, we do not have the notion of a population (or a species) in AI. But this can be taken as an opportunity rather than a limitation, and think of notions of difficulty that are not based on a population. The advantage in AI is that tasks can be understood computationally, and difficulty can be linked to several theoretical and empirical notions of complexity and resources in the field.

Overall, the vindication of generality as a standalone score reframes the question of what "General Intelligence" is and how it can be measured in a different way, disentangling the conflation between generality and capability. For machines, it can help recover the meaning that the G in "Artificial General Intelligence" was originally meant to have.

The finally message of this report goes clearly in the direction of future work. First, there is much to do to further clarify existing and newly-introduced notions of difficulty for different kinds of tasks, or even universally, for all tasks, and use them in the analysis of results. Second, using ad-hoc measures of difficulty (even if they are imperfect), we can already analyse the individual generality of a myriad of results already collected for humans and non-human animals, and an increasing number of repositories of results for AI systems.

Acknowledgements

I thank David Stillwell and Aiden Loe for suggesting the relation between person-fit metrics and an early version of generality, and Heinrich Peters for pointing out Guttman's model. I also thank Fernando Martínez-Plumed for several discussions on the idea of generality as the slope of the ACC derived from a logistic model and how this can be obtained from AI benchmarks.

This work has been partially supported by the EU (FEDER) and Spanish MINECO grant TIN2015-69175-C4-1-R, and by Generalitat Valenciana PROMETEOII/2015/013. Part of this work has been done while visiting the Leverhulme Centre for the Future of Intelligence, generously funded by the Leverhulme Trust. I also thank the UPV for granting me a sabbatical leave and the funding from the Spanish MECD programme "Salvador de Madariaga" (PRX17/00467) and the Generalitat Valenciana grants for research stays.

Appendix A. Proofs

In this appendix we include the proofs of the properties, lemmata and propositions.

Properties of generality

Despite being straightforward, in what follows we include the proofs for the properties presented in section 3.3.

Proposition 4. *Given an agent with capability ψ_j, any positive translation by k implies that capability becomes $\psi_j + k$.*

Proof. A translation creates a new function such that $h' \leftarrow h - k$ and $\psi_j^{[h']} = 1$ for all $h' < k$, so the new capability ψ'_j is now:

$$\psi'_j \quad = \quad \int_0^k 1\, dh' + \int_k^\infty \psi_j^{[h-k]}\, dh = k + \psi_j$$

□

Proposition 5. *Given an agent with capability ψ_j where the $|l|$-leftmost part of the original curve was saturated, any negative translation by $k \leqslant l$ implies that capability becomes $\psi_j - k$.*

Proof. As the left part of the curve is saturated, ψ_j can be decomposed into

$$\psi_j \quad = \quad \int_0^l 1\, dh + \int_l^\infty \psi_j^{[h]}\, dh$$

Now the translation removes part of the first term, so the new capability is:

$$\psi'_j \quad = \quad \int_0^{l-k} 1\, dh + \int_{l-k}^\infty \psi_j^{[h+l-k]}\, dh = l - k + \psi_j - l = \psi_j - k$$

□

Proposition 6. *With the same conditions as the above two propositions, generality is invariant to translation.*

Proof. For a positive translation, we have that the new effort m'_j equals:

$$\begin{aligned} m'_j &= \int_0^k h' \cdot 1\, dh' + \int_k^\infty h\psi_j^{[h-k]}\, dh \\ &= \frac{k^2}{2} + \int_k^\infty (h-k)\psi_j^{[h-k]}\, dh + \int_k^\infty k\psi_j^{[h-k]}\, dh \\ &= \frac{k^2}{2} + m_j + k\psi_j \end{aligned}$$

From proposition 4 we have that $\psi'_j = k + \psi_j$. Putting both things together into the definition of spread (Eq. 4), we have:

$$\begin{aligned} z'_j &= \sqrt{2m'_j - \psi_j'^2} = \sqrt{2\frac{k^2}{2} + 2m_j + 2k\psi_j - (k+\psi_j)^2} \\ &= \sqrt{k^2 + 2m_j + 2k\psi_j - k^2 - 2k\psi_j - \psi_j^2} = \sqrt{2m_j - \psi_j^2} \end{aligned}$$

As generality is the reciprocal of spread, and spread does not change, then it is invariant to positive translation. The proof for the negative translation is similar. □

Proposition 7. *Compactness: any mass moved to the left of the plot such that* $\psi_j^{[h_1:h_2]} \leftarrow \psi_j^{[h_1:h_2]} + q$ *while* $\psi_j^{[h_3:h_4]} \leftarrow \psi_j^{[h_3:h_4]} - q$*, with* $h_2 < h_3$ *will increase* γ_j*.*

Proof. Clearly, $\psi'_j = \psi_j$, since the same mass q is included in the integral one way or the other. We have that the new effort m'_j:

$$m'_j \quad = \quad \int_0^{h_1} h\psi_j^{[h]}\, dh + \int_{h_1}^{h_2} h\psi_j'^{[h]}\, dh + \int_{h_2}^{h_3} h\psi_j^{[h]}\, dh + \int_{h_3}^{h_4} h\psi_j'^{[h]}\, dh + \int_{h_4}^{\infty} h\psi_j^{[h]}\, dh$$

Since $h_2 < h_3$, we have that $\int_{h_1}^{h_2} h\psi_j'^{[h]}\, dh + \int_{h_3}^{h_4} h\psi_j'^{[h]}\, dh < \int_{h_1}^{h_2} h\psi_j^{[h]}\, dh + \int_{h_3}^{h_4} h\psi_j^{[h]}\, dh$, and hence $m'_j < m_j$. Now, from the definition of spread (Eq. 4), we have:

$$z'_j \quad = \quad \sqrt{2m'_j - \psi_j'^2} = \sqrt{2m'_j - \psi_j^2} < \sqrt{2m_j - \psi_j^2} = z_j$$

As spread is smaller, and generality is the reciprocal, this completes the proof. □

Corollary 8. *Maximum: given a fixed capability* ψ_j*, the minimum expected difficulty* $\mathbb{H}_j$ *and the maximum generality* γ_j *are obtained with a step agent characteristic function on* $h = \psi_j$*.*

Proof. By proposition 7, generality is increased as far as we move mass of the function from right to left, while keeping the area constant. This means that the maximum area with highest generality is obtained by a step function, whose location must be on $h = \psi_j$. □

Proposition 9. *Step function: given a step function, capability is double the expected difficulty (i.e.,* $\psi_j = 2\mathbb{H}_j$*), and generality* $\gamma_j = \infty$*.*

Proof. The area of a step function with location l is:

$$\psi_j \quad = \quad \int_0^{l} 1\, dh = l$$

As effort in this situation is:

$$m_j \quad = \quad \int_0^{l} h \cdot 1\, dh = \frac{l^2}{2}$$

Expected difficulty is just:

$$\mathbb{H}_j \quad = \quad \frac{m_j}{\psi_j} = \frac{l^2}{2} = \frac{l}{2}$$

So $\psi_j = 2\mathbb{H}_j$ and $z_j = \sqrt{2\frac{l^2}{2} - l^2} = 0$, so its reciprocal is ∞. □

Proposition 10. *Constant: given a constant function $\psi_j^{[h]} = c$ from 0 to q, we have $\psi_j = cq$, $z_j = \sqrt{c(1-c)}q$ and $\gamma_j = 1/\sqrt{c(1-c)}q$.*

Proof. We have:

$$\psi_j \quad = \quad \int_0^q c\, dh = cq$$

and

$$m_j \quad = \quad \int_0^q h \cdot c\, dh = c\frac{q^2}{2}$$

and

$$z_j \quad = \quad \sqrt{2m_j - \psi_j^2} = \sqrt{2c\frac{q^2}{2} - (cq)^2} = \sqrt{c(1-c)q^2} = \sqrt{c(1-c)}q$$

□

We have two examples at the bottom of Figure 2. For instance, on the left we have $\psi_j = cq = 0.2 \cdot 2 = 5 \approx 4.97$, $z_j = \sqrt{c(1-c)}q = \sqrt{0.2(0.8)}25 = 10 \approx 9.93$ and $\gamma_j = 1/\sqrt{c(1-c)}q = 1/10 \approx 0.10$. The precision divergence is given because the curves are not perfectly flat.

Proposition 11. *Task transitivity: if an agent π_j is s-saturated then for every task μ_b such that $A_b^j = 1$ in the saturated area then for all other tasks a with $\hbar(\mu_a) \leqslant \hbar(\mu_b)$ we have that $A_a^j = 1$.*

Proof. If an agent s is s-saturated then $A_b^j = 1$ for all tasks such that $\hbar(\mu_b) \leqslant s$. If b is in the saturated area, any other task a of lower difficulty also is. □

Proposition 12. *Agent transitivity: if two agents π_a and π_b have maximum generality $\gamma_a = \gamma_b = \infty$ and $\psi_a \leqslant \psi_b$ then for every task μ_i such that $A_i^a = 1$ then $A_i^b = 1$.*

Proof. It is sufficient to see that both agents will have step functions. □

Note that if the generality of π_b is not infinite, it is not sufficient to have a curve for π_b that covers the curve for π_a. The reason is that there might be values of h for which $0 < \psi_a^{[h]} < \psi_b^{[h]} < 1$, and in these cases some of tasks that make the non-zero value in $\psi_a^{[h]}$ might not be in the tasks that make the value of $\psi_b^{[h]}$.

Proposition 13. *Same units: if we introduce a unit for difficulty, let us call it bints (for basic intelligence units), then capability is also measured in bints, spread is also measured in bints and hence generality is measured in $1/bints$.*

Proof. As ψ_j is an integral over difficulty and the domain of the function is unitless (accomplishment, which is a proportion), then ψ_j has the same units as difficulty. As m_j includes the factor h in the integral, i.e.,

$$m_j \quad = \quad \int_k^\infty h\psi_j^{[h]}\, dh$$

the result is in $bints^2$. Finally, from the definition of spread:

$$z_j \quad = \quad \sqrt{2m_j - \psi_j^2}$$

we get $\sqrt{bints^2}$, which means that spread is measured in *bints*, and the reciprocal for generality. □

Proofs when using a normal distribution for capability

Here we include the proofs about the case where the ACC derives from assuming a normal distribution on capability.

Proposition 14. *(proposition 1 in the paper) Assuming a normal distribution on capability, with standard deviation σ, the slope of the ACC will be $-\frac{1}{\sigma\sqrt{2\pi}}$.*

Proof. (of proposition 1) We know that a normal distribution with mean μ and standard deviation σ will lead to the following agent characteristic curve:

$$\psi_j^{[h]} \quad = \quad 1 - \Phi\left(\frac{h-\mu}{\sigma}\right)$$

with Φ being the CDF of the standard normal distribution and ϕ being the density function of the standard normal distribution. The maximum slope of this is the first derivative at μ, which is:

$$\text{slope} \quad = \quad -\phi\left(\frac{\mu-\mu}{\sigma}\right) = -\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(0)^2}{2\sigma^2}} = -\frac{1}{\sqrt{2\pi\sigma^2}}$$

and hence the slope of the ACC will be $-\frac{1}{\sigma\sqrt{2\pi}}$. □

Lemma 15. *(lemma 2 in the paper) Assuming a normal distribution on the capability, with mean μ and standard deviation σ, such that the location is sufficiently beyond 0 to have negligible mass below 0 (i.e., $\frac{\mu}{\sigma} \gg 0$), we have that $m_j = \frac{\sigma^2+\mu^2}{2}$.*

Proof. (of lemma 2) As in proposition 1), we know that a normal distribution with mean μ and standard deviation σ will lead to the following agent characteristic curve:

$$\psi_j^{[h]} \quad = \quad 1 - \Phi\left(\frac{h-\mu}{\sigma}\right)$$

We plug this into the definition of effort and operate a little bit on it in order to put the expression in terms of the cumulative distribution function Φ of the normal distribution:

$$\begin{aligned} m_j &= \int_0^\infty h \cdot \left(1 - \Phi\left(\frac{h-\mu}{\sigma}\right)\right) dh \\ &= -\int_{-\infty}^0 h \cdot \Phi\left(\frac{h+\mu}{\sigma}\right) dh \end{aligned}$$

Fortunately, we can find the following integral of the moment of the CDF on page 402 (second last, entry 10,001) in [98]:

$$\int x\Phi(a+bx)\,dx = \frac{1}{2b^2}\left((b^2x^2 - a^2 - 1)\Phi(a+bx) + (bx-a)\phi(a+bx)\right) + C$$

And ϕ is the density function.

In our case, $a = \frac{\mu}{\sigma}$ and $b = \frac{1}{\sigma}$, so we can put all things together into:

$$\begin{aligned} m_j &= -\left[\frac{1}{2b^2}\left((b^2x^2 - a^2 - 1)\Phi(a+bx) + (bx-a)\phi(a+bx)\right)\right]_{-\infty}^0 \\ &= -\left[\frac{1}{2b^2}\left((-a^2-1)\Phi(a) - a\phi(a)\right)\right] - [0+0]] \\ &= \frac{1}{2\left(\frac{1}{\sigma}\right)^2}\left(\left(\left(\frac{\mu}{\sigma}\right)^2 + 1\right)\Phi\left(\frac{\mu}{\sigma}\right) + \frac{\mu}{\sigma}\phi\left(\frac{\mu}{\sigma}\right)\right) \end{aligned}$$

Since we are assuming that $\frac{\mu}{\sigma} \gg 0$, we have that $\Phi(\frac{\mu}{\sigma}) \approx 1$ and $\phi(\frac{\mu}{\sigma}) \approx 0$, so we get:

$$\begin{aligned} m_j &= \frac{1}{2(\frac{1}{\sigma})^2}\left(\left(\frac{\mu}{\sigma}\right)^2 + 1\right) = \frac{\sigma^2\frac{\mu^2}{\sigma^2} + \sigma^2}{2} \\ &= \frac{\mu^2 + \sigma^2}{2} \end{aligned}$$

□

Proposition 16. *(proposition 3 in the paper) With the same assumptions as lemma 2, we have that spread $z_j = \sigma$ and $\gamma = \frac{1}{\sigma}$.*

Proof. (of proposition 3) As the normal distribution is symmetric, we have that the location of the CDF is of course μ, so the capability $\psi_j = \mu$, and plugging m_j from lemma 2, we have:

$$\begin{aligned} z_j &= \sqrt{2m_j - \psi_j^2} = \sqrt{2m_j - \mu^2} \\ &= \sqrt{2\left(\frac{\mu^2+\sigma^2}{2}\right) - \mu^2} = \sigma \end{aligned}$$

And by the definition of generality we have $\gamma_j = \frac{1}{\sigma}$. □

References

[1] Sam Adams, Itmar Arel, Joscha Bach, Robert Coop, Rod Furlan, Ben Goertzel, J Storrs Hall, Alexei Samsonovich, Matthias Scheutz, Matthew Schlesinger, Stuart C. Shapiro, and John Sowa. Mapping the landscape of human-level artificial general intelligence. *AI magazine*, 33(1):25–42, 2012. 3, 27

[2] John Allman, Todd McLaughlin, and Atiya Hakeem. Brain weight and life-span in primate species. *Proceedings of the National Academy of Sciences*, 90(1):118–122, 1993. 21

[3] Britt Anderson. The g factor in non-human animals. In Jamie A. Goode Gregory R. Bock and Kate Webb, editors, *The nature of intelligence*, volume 233, pages 79–95. John Wiley & Sons, 2000. 3

[4] Haris Aziz, Markus Brill, Felix Fischer, Paul Harrenstein, Jérôme Lang, and Hans Georg Seedig. Possible and necessary winners of partial tournaments. *Journal of Artificial Intelligence Research*, 54:493–534, 2015. 35

[5] Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. On the effect of calibration in classifier combination. *Applied intelligence*, 38(4):566–585, 2013. 19

[6] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 06 2013. 34

[7] Tarek Besold, José Hernández-Orallo, and Ute Schmid. Can machine intelligence be measured in the same way as human intelligence? *KI-Künstliche Intelligenz*, 29(3):291–297, 2015. 32

[8] Tarek R Besold and Ute Schmid. Why generality is key to human-level artificial intelligence. *Advances in Cognitive Systems*, (4):13–24, 2016. 28

[9] Philip Bontrager, Ahmed Khalifa, Andre Mendes, and Julian Togelius. Matching games and algorithms for general video game playing. In *Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference*, pages 122–128, 2016. 4

[10] Judith M Burkart, Michèle N Schubiger, and Carel P van Schaik. The evolution of general intelligence. *Behavioral and Brain Sciences*, 40, 2017. 3, 20, 21, 23

[11] Richard Byrne and Andrew Whiten. Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans (oxford science publications). 1989. 19

[12] Davide Castelvecchi. Tech giants open virtual worlds to bevy of AI programs. *Nature*, 540:323–324, 2016. 34

[13] Rafael Jaime De Ayala. *Theory and practice of item response theory*. Guilford Publications, 2009. 5, 11, 12

[14] I. J. Deary, V. Egan, G. J. Gibson, E. J. Austin, C. R. Brand, and T. Kellaghan. Intelligence and the differentiation hypothesis. *Intelligence*, 23(2):105–132, 1996. 17, 18

[15] D. K. Detterman. General intelligence: Cognitive and biological explanations. In R. J. Sternberg and E. L. Grigorenko, editors, *The general factor of intelligence: How general is it?*, pages 223–243. 2002. 3

[16] D. K. Detterman and M. H. Daniel. Correlations of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence*, 13(4):349–359, 1989. 17, 18

[17] C. Dimitrakakis, G. Li, and N. Tziortziotis. The reinforcement learning competition 2014. *AI Magazine*, 35(3):61–65, 2014. 34

[18] D. L. Dowe and J. Hernandez-Orallo. IQ tests are not for machines, yet. *Intelligence*, 40(2):77–81, 2012. 32

[19] D. L. Dowe, J. Hernández-Orallo, and P. K. Das. Compression and intelligence: social environments and communication. In J. Schmidhuber, K.R. Thórisson, and M. Looks, editors, *Artificial General Intelligence*, volume 6830, pages 204–211. LNAI series, Springer, 2011. 36

[20] David L Dowe and José Hernández-Orallo. How universal can an intelligence test be? *Adaptive Behavior*, 22(1):51–69, 2014. 33

[21] B. Edmonds. The social embedding of intelligence. In Robert Epstein, Gary Roberts, and Grace Beber, editors, *Parsing the Turing Test*, pages 211–235. Springer, 2009. 29

[22] A. E. Elo. *The rating of chessplayers, past and present*, volume 3. Batsford London, 1978. 35

[23] S. E. Embretson and S. P. Reise. *Item response theory for psychologists*. L. Erlbaum, 2000. 5, 11, 12

[24] George W Ernst and Allen Newell. *GPS: A case study in generality and problem solving*. Academic Press, 1969. 26

[25] Pere J Ferrando. A general approach for assessing person fit and person reliability in typical-response measurement. *Applied Psychological Measurement*, 38(2):166–183, 2014. 13, 14

[26] Jerry A Fodor. *The modularity of mind: An essay on faculty psychology*. MIT press, 1983. 19

[27] G. J. Fogarty and L. Stankov. Challenging the "law of diminishing returns". *Intelligence*, 21(2):157–174, 1995. 18

[28] M. Genesereth, N. Love, and B. Pell. General game playing: Overview of the AAAI competition. *AI Magazine*, 26(2):62, 2005. 34

[29] M. Genesereth and M. Thielscher. General game playing. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(2):1–229, 2014. 34

[30] Louis Guttman. A basis for scaling qualitative data. *American sociological review*, 9(2):139–150, 1944. 10, 12

[31] Louis Guttman. The basis for scalogram analysis. In Samuel A Stouffer, Louis Guttman, Edward A Suchman, Paul F Lazarsfeld, Shirley A Star, and John A Clausen, editors, *Measurement and prediction*, pages 60–90. Princeton University Press, 1950. 10, 12

[32] Louis Guttman and Ruth Guttman. A theory of behavioral generality and specificity during mild stress. *Systems Research and Behavioral Science*, 21(6):469–477, 1976. 15

[33] J. Hernández-Orallo. Beyond the Turing Test. *J. Logic, Language & Information*, 9(4):447–466, 2000. 28, 29, 30, 31

[34] J. Hernández-Orallo. Computational measures of information gain and reinforcement in inference processes. *AI Communications*, 13(1):49–50, 2000. 36

[35] J. Hernández-Orallo. On the computational measurement of intelligence factors. In A. M. Meystel and E. R. Messina, editors, *Measuring the performance and intelligence of systems: proceedings of the 2000 PerMIS Workshop, August 14–16, 2000*, pages 72–79. National Institute of Standards and Technology (NIST) Special Publication 970, Gaithersburg, MD, U.S.A., 2000. 29

[36] J. Hernández-Orallo. A (hopefully) non-biased universal environment class for measuring intelligence of biological and artificial systems. In M. Hutter et al., editor, *Artificial General Intelligence, 3rd Intl Conf*, pages 182–183. Atlantis Press, 2010. Extended report at http://users.dsic.upv.es/proy/anynt/unbiased.pdf. 33

[37] J. Hernández-Orallo. C-tests revisited: Back and forth with complexity. In J. Bieger, B. Goertzel, and A. Potapov, editors, *Artificial General Intelligence - 8th International Conference, AGI 2015, Berlin, Germany, July 22-25, 2015, Proceedings*, pages 272–282. Springer, 2015. 29, 30

[38] J. Hernández-Orallo. A note about the generalisation of the c-tests. *arXiv preprint, arXiv:1412.8529*, 2015. 29, 30

[39] J. Hernández-Orallo. Stochastic tasks: Difficulty and Levin search. In J. Bieger, B. Goertzel, and A. Potapov, editors, *Artificial General Intelligence - 8th International Conference, AGI 2015, Berlin, Germany, July 22-25, 2015, Proceedings*, pages 90–100. Springer, 2015. 31

[40] J. Hernández-Orallo. Is Spearman's law of diminishing returns (SLODR) meaningful for artificial agents? In *ECAI 2016 - 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*, pages 471–479, 2016. 19

[41] J. Hernández-Orallo. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press, 2017. 3, 6, 28, 29, 30, 31, 32, 37

[42] J. Hernández-Orallo and D. L. Dowe. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18):1508 – 1539, 2010. 29, 32

[43] J. Hernández-Orallo, D. L. Dowe, S. España-Cubillo, M. V. Hernández-Lloreda, and J. Insa-Cabrera. On more realistic environment distributions for defining, evaluating and developing intelligence. In J. Schmidhuber, K.R. Thórisson, and M. Looks, editors, *Artificial General Intelligence*, volume 6830, pages 82–91. LNAI, Springer, 2011. 35

[44] J. Hernández-Orallo, D. L. Dowe, and M. V. Hernández-Lloreda. Universal psychometrics: Measuring cognitive abilities in the machine kingdom. *Cognitive Systems Research*, 27:5074, 2014. 33, 37

[45] J. Hernández-Orallo, D. L. Dowe, and M. V. Hernández-Lloreda. Measuring cognitive abilities of machines, humans and non-human animals in a unified way: towards universal psychometrics. *Technical Report 2012/267, Faculty of Information Technology, Clayton School of I.T., Monash University, Australia*, March 2012. 33

[46] J. Hernández-Orallo and I. García-Varea. Explanatory and creative alternatives to the MDL principle. *Foundations of Science*, 5(2):185–207, 2000. 36

[47] J. Hernández-Orallo, J. Insa-Cabrera, D. L. Dowe, and B. Hibbard. Turing machines and recursive Turing Tests. In V. Muller and A. Ayesh, editors, *AISB/IACAP 2012 Symposium "Revisiting Turing and his Test"*, pages 28–33. The Society for the Study of Artificial Intelligence and the Simulation of Behaviour, 2012. 36

[48] José Hernández-Orallo. A computational definition of consilience. *Philosophica*, 61(1):19–37, 1998. 36

[49] José Hernández-Orallo. Constructive reinforcement learning. *International Journal of Intelligent Systems*, 15(3):241–264, 2000. 36

[50] José Hernández-Orallo. Universal psychometrics tasks: difficulty, composition and decomposition. *arXiv preprint arXiv:1503.07587*, 2015. 31

[51] José Hernández-Orallo. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, 48(3):397–447, 2017. 34

[52] José Hernández-Orallo, Marco Baroni, Jordi Bieger, Nader Chmait, David L Dowe, Katja Hofmann, Fernando Martínez-Plumed, Claes Strannegård, and Kristinn R Thórisson. A new AI evaluation cosmos: Ready to play the game? *AI Magazine*, 38(3), 2017. 34

[53] José Hernández-Orallo and David L Dowe. On potential cognitive abilities in the machine kingdom. *Minds and Machines*, 23(2):179–210, 2013. 35

[54] José Hernández-Orallo, Javier Insa-Cabrera, David L Dowe, and Bill Hibbard. Turing tests with turing machines. *Turing-100*, 10:140–156, 2012. 36

[55] José Hernández-Orallo, Fernando Martínez-Plumed, Ute Schmid, Michael Siebers, and David L Dowe. Computer models solving intelligence test problems: Progress and implications. *Artificial Intelligence*, 230:74–107, 2016. 32

[56] José Hernández-Orallo and Neus Minaya-Collado. A formal definition of intelligence based on an intensional variant of algorithmic complexity. In *Proceedings of International Symposium of Engineering of Intelligent Systems (EIS98)*, pages 146–163, 1998. 30, 31

[57] José Hernández-Orallo and M José Ramírez-Quintana. Predictive software. *Automated Software Engineering*, 8(2):139–166, 2001. 36

[58] Esther Herrmann, Josep Call, María Victoria Hernández-Lloreda, Brian Hare, and Michael Tomasello. Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *science*, 317(5843):1360–1366, 2007. 21

[59] B. Hibbard. Bias and no free lunch in formal measures of intelligence. *Journal of Artificial General Intelligence*, 1(1):54–61, 2009. 29, 32

[60] Bill Hibbard. Adversarial sequence prediction. *Frontiers in Artificial Intelligence and Applications*, 171:399, 2008. 36

[61] Douglas R Hofstadter. *Gödel, escher, bach*. Vintage Books New York, 1980. 34

[62] Kay E Holekamp. Questioning the social intelligence hypothesis. *Trends in cognitive sciences*, 11(2):65–69, 2007. 19

[63] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, 2005. 28

[64] M. Hutter. Universal algorithmic intelligence: A mathematical top→down approach. In B. Goertzel and C. Pennachin, editors, *Artificial General Intelligence*, Cognitive Technologies, pages 227–290. Springer, Berlin, 2007. 37

[65] Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2004. 37

[66] C. Igel and M. Toussaint. A no-free-lunch theorem for non-uniform distributions of target functions. *Journal of Mathematical Modelling and Algorithms*, 3(4):313–322, 2005. 28

[67] J. Insa-Cabrera, D. L. Dowe, and J. Hernández-Orallo. Evaluating a reinforcement learning algorithm with a general intelligence test. In J.A. Lozano, J.A. Gamez, and J.A. Moreno, editors, *Current Topics in Artificial Intelligence. CAEPIA 2011*. LNAI Series 7023, Springer, 2011. 33

[68] Javier Insa-Cabrera, José-Luis Benacloch-Ayuso, and José Hernández-Orallo. On measuring social intelligence: Experiments on competition and cooperation. In *International Conference on Artificial General Intelligence*, pages 126–135. Springer, 2012. 35

[69] Javier Insa-Cabrera, David L Dowe, Sergio España-Cubillo, M Victoria Hernández-Lloreda, and José Hernández-Orallo. Comparing humans and ai agents. In *International Conference on Artificial General Intelligence*, pages 122–132. Springer, 2011. 33

[70] Javier Insa-Cabrera and José Hernández-Orallo. Instrumental properties of social testbeds. In *International Conference on Artificial General Intelligence*, pages 101–110. Springer, 2015. 35

[71] Javier Insa-Cabrera, José Hernández-Orallo, David L Dowe, Sergio Espana, and M Victoria Hernández-Lloreda. The anynt project intelligence test: Lambda-one. In *AISB/IACAP 2012 Symposium Revisiting Turing and his Test*, pages 20–27, 2012. 29

[72] Ivo Jacobs and Peter Gärdenfors. The false dichotomy of domain-specific versus domain-general cognition. *Behavioral and Brain Sciences*, 40, 2017. 22

[73] A. R. Jensen. *The g factor: The science of mental ability*. Westport, Praeger, 1998. 3, 14, 18

[74] Alison Jolly. Lemur social behavior and primate intelligence. *Science*, 153(3735):501–506, 1966. 19

[75] Michael T Kane and Robert L Brennan. Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, 4(1):105–126, 1980. 12

[76] Kenneth E Kinnear Jr. Generality and difficulty in genetic programming: Evolving a sort. In *ICGA*, pages 287–294. Citeseer, 1993. 36

[77] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2004. 19

[78] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003. 4, 19

[79] James I Lathrop. Compression depth and genetic programs. *Genetic Programming*, pages 370–379, 1997. 36

[80] Louis Lefebvre. Brains, innovations, tools and cultural transmission in birds, non-human primates, and fossil hominins. *Frontiers in human neuroscience*, 7:245, 2013. 21

[81] S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007. 29, 30, 32

[82] L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9(3):265–266, 1973. 31

[83] Leonid A Levin. Universal heuristics: How do humans solve unsolvable problems? In D. L. Dowe, editor, *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*, volume 7070 of *Lecture Notes in Computer Science*, pages 53–54. Springer, 2013. 31

[84] M. Li and P. Vitányi. *An introduction to Kolmogorov complexity and its applications (3rd ed.)*. Springer-Verlag, 2008. 32

[85] James Lumsden. Person reliability. *Applied Psychological Measurement*, 1(4):477–482, 1977. 13

[86] Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. arXiv preprint arXiv:1709.06009, 2017. 34

[87] Fernando Martínez-Plumed, Cesar Ferri, José Hernández-Orallo, and María J Ramírez-Quintana. Knowledge acquisition with forgetting: an incremental and developmental setting. *Adaptive Behavior*, 23(5):283–299, 2015. 35

[88] Fernando Martínez-Plumed, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. A computational analysis of general intelligence tests for evaluating cognitive development. *Cognitive Systems Research*, 43:100–118, 2017. 32

[89] Fernando Martınez-Plumed and José Hernández-Orallo. Ai results for the atari 2600 games: difficulty and discrimination using irt. *EGPAI, Evaluating General-Purpose Artificial Intelligence*, 2016. 33

[90] Fernando Martínez-Plumed, Ricardo B. C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Making sense of item response theory in machine learning. In *ECAI 2016 - 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*, pages 1140–1148, 2016. 33

[91] John McCarthy. Generality in artificial intelligence. *Communications of the ACM*, 30(12):1030–1035, 1987. 3, 27, 36

[92] J. McDermott, D. R. White, S. Luke, L. Manzoni, M. Castelli, L. Vanneschi, W. Jaśkowski, K. Krawiec, R. Harper, K. De Jong, and U.-M. O'Reilly. Genetic programming needs better benchmarks. In *Proceedings of the 14th international conference on genetic and evolutionary computation conference*, pages 791–798. ACM, 2012. 34

[93] Rob R Meijer and Klaas Sijtsma. Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2):107–135, 2001. 11

[94] Andre Mendes, A Nealen, and J Togelius. Hyperheuristic general video game playing. In *Proceedings of Computational Intelligence and Games (CIG). IEEE*, 2016. 4

[95] Robert J Mokken and Charles Lewis. A nonparametric approach to the analysis of dichotomous item responses. *Applied psychological measurement*, 6(4):417–430, 1982. 12

[96] Allen Newell, John C Shaw, and Herbert A Simon. Report on a general problem-solving program. In *IFIP Congress*, pages 256–264, 1959. 3, 26

[97] Thorbjrn S. Nielsen, Gabriella AB Barros, Julian Togelius, and Mark J Nelson. Towards generating arcade game rules with VGDL. In *Computational Intelligence and Games (CIG), 2015 IEEE Conference on*, pages 185–192. IEEE, 2015. 4

[98] Donald Bruce Owen. A table of normal integrals. *Communications in Statistics-Simulation and Computation*, 9(4):389–419, 1980. 43

[99] D. Perez, S. Samothrakis, J. Togelius, T. Schaul, S. Lucas, A. Couëtoux, J.l Lee, C.-U. Lim, and T. Thompson. The 2014 general video game playing competition. *IEEE Transactions on Computational Intelligence and AI in Games*, 8:229–243, 2015. 34

[100] Jean Piaget and Margaret Cook. *The origins of intelligence in children*, volume 8. International Universities Press New York, 1952. 19

[101] Ricardo BC Prudêncio, José Hernández-Orallo, and Adolfo Martınez-Usó. Analysis of instance hardness in machine learning using item response theory. In *Second International Workshop on Learning over Multiple Contexts in ECML 2015. Porto, Portugal, 11 September 2015*, volume 1, 2015. 33

[102] T. Schaul. An extensible description language for video games. *Computational Intelligence and AI in Games, IEEE Transactions on*, 6(4):325–331, 2014. 34

[103] Amanda Seed, Nathan Emery, and Nicola Clayton. Intelligence in corvids and apes: a case of convergent evolution? *Ethology*, 115(5):401–420, 2009. 19

[104] David M Shuker, Louise Barrett, Thomas E Dickins, Thom C Scott-Phillips, and Robert A Barton. General intelligence does not help us understand cognitive evolution. *Behavioral and Brain Sciences*, 40, 2017. 25

[105] Klaas Sijtsma and Ivo W Molenaar. *Introduction to nonparametric item response theory*, volume 5. Sage, 2002. 12

[106] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017. 27, 35

[107] Daniel Sol, Simon Ducatez, and Ferran Sayol. Cognitive buffer hypothesis, the. *Encyclopedia of Evolutionary Psychological Science*, pages 1–6, 2016. 21

[108] R. J. Solomonoff. A preliminary report on a general theory of inductive inference, 1960. Report V-131, Zator Co., Cambridge, Ma. Feb 4, revision, Nov. 28

[109] R. J. Solomonoff. A formal theory of inductive inference. Part I. *Information and control*, 7(1):1–22, 1964. 28

[110] C. Spearman. General Intelligence, Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–92, 1904. 3, 14

[111] C. Spearman. *The abilities of man: Their nature and measurement*. Macmillan, New York, 1927. 14, 18

[112] R. J. Sternberg. *Handbook of intelligence*. Cambridge University Press, 2000. 14

[113] Kikumi K Tatsuoka and Maurice M Tatsuoka. Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, 20(3):221–230, 1983. 12

[114] Tom E Trabin and David J Weiss. The person response curve: Fit of individuals to item characteristic curve models. Technical report, Minnesota Univ. Minneapolis Dept. of Psychology, 1979. 11

[115] Tom E Trabin and David J Weiss. The person response curve: Fit of individuals to item response theory models. In *New horizons in testing*, pages 83–108. Elsevier, 1983. 11

[116] E. M. Tucker-Drob. Differentiation of cognitive abilities across the life span. *Developmental psychology*, 45(4):1097, 2009. 17

[117] C David Vale and David J Weiss. A study of computer-administered stradaptive ability testing. Technical report, Minnesota Univ. Minneapolis Dept. of Psychology, 1975. 11

[118] Jayden O van Horik and Stephen EG Lea. Disentangling learning from knowing: Does associative learning ability underlie performances on cognitive test batteries? *Behavioral and Brain Sciences*, 40, 2017. 31

[119] Carel P Van Schaik, Karin Isler, and Judith M Burkart. Explaining brain size variation: from social to cultural brain. *Trends in cognitive sciences*, 16(5):277–284, 2012. 21

[120] Paul Vitányi and Ming Li. On prediction by data compression. In *Machine Learning: ECML-97*, pages 14–30. Springer, 1997. 36

[121] D. R. White, J. McDermott, M. Castelli, L. Manzoni, B. W. Goldman, G. Kronberger, W. Jaśkowski, U.-M. O'Reilly, and S. Luke. Better GP benchmarks: Community survey results and proposals. *Genetic Programming and Evolvable Machines*, 14:3–29, 2013. 34

[122] Andrew Whiten and Richard W Byrne. *Machiavellian intelligence II: Extensions and evaluations*, volume 2. Cambridge University Press, 1997. 19

[123] S. Whiteson, B. Tanner, and A. White. The Reinforcement Learning Competitions. *The AI magazine*, 31(2):81–94, 2010. 34

[124] D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996. 28

[125] D. H. Wolpert. What the no free lunch theorems really mean; how to improve search algorithms. Technical report, Santa fe Institute Working Paper, 2012. 28

[126] D. H. Wolpert and W. G. Macready. No free lunch theorems for search. Technical report, SFI-TR-95-02-010 (Santa Fe Institute), 1995. 28

[127] Michael A Woodley and Edward Bell. Is collective intelligence (mostly) the general factor of personality? a comment on Woolley, Chabris, Pentland, Hashmi and Malone (2010). *Intelligence*, 39(2):79–81, 2011. 19

[128] Michael A. Woodley of Menie, Heitor BF Fernandes, Jan te Nijenhuis, Mateo Peñaherrera-Aguirre, and Aurelio José Figueredo. General intelligence is a source of individual differences between species: Solving an anomaly. *Behavioral and Brain Sciences*, 40, 2017. 24, 26

[129] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688, 2010. 19