Universal Psychometrics:
Measuring Cognitive Abilities in the Machine Kingdom

José Hernández-Orallo[a,*], David L. Dowe[b], M.Victoria Hernández-Lloreda[c]

[a]*DSIC, Universitat Politècnica de València, València, Spain.*
jorallo@dsic.upv.es
[b]*Computer Science and Software Engineering, Clayton School of Information Technology Monash University, Clayton, Vic. 3800, Australia.*
david.dowe@infotech.monash.edu.au
[c]*Departamento de Metodología de las Ciencias del Comportamiento, Universidad Complutense de Madrid, Spain.*
vhlloreda@psi.ucm.es

Abstract

We present and develop the notion of 'universal psychometrics' as a subject of study, and eventually a discipline, that focusses on the measurement of cognitive abilities for *the machine kingdom*, which comprises any (*cognitive*) *system*, individual or collective, either artificial, biological or hybrid. Universal psychometrics can be built, of course, upon the experience, techniques and methodologies from (human) psychometrics, comparative cognition and related areas. Conversely, the perspective and techniques which are being developed in the area of machine intelligence measurement using (algorithmic) information theory can be of much broader applicability and implication outside artificial intelligence. This general approach to universal psychometrics spurs the re-understanding of most (if not all) of the big issues about the measurement of cognitive abilities, and creates a new foundation for (re)defining and mathematically formalising the concept of cognitive task, evaluable subject, interface, task choice, difficulty, agent response curves, etc. We introduce the notion of a universal cognitive test and discuss whether (and when) it may be necessary for exploring the machine kingdom. On the issue of intelligence and very general abilities, we also get some results and connections with the related notions of no-free-lunch theorems and universal priors.

Keywords: cognitive system evaluation; psychometrics; cognitive abilities; artificial intelligence; algorithmic information theory

1. Introduction

The measurement of intelligence and other cognitive abilities is the subject of several disciplines. On one hand, psychometrics has developed valid and reliable tests for *human* intelligence and related abilities. On the other hand, comparative psychology (and cognition) has explored *animal* behaviour and the range of cognitive abilities for many different species.

---

*Corresponding author. Address: "DSIC, Universitat Politècnica de València, Camí de Vera s/n, 46022 València, Spain", Phone: +34 96 3877007 Ext.: 73585, Fax: +34 96 3877359

While psychometrics and comparative psychology usually work independently, there are areas and approaches where their connection becomes more explicty, such as evolutionary psychology, the analysis of the personality of animals (Gosling, 2001) and the studies of comparative cognition on apes (Herrmann et al., 2007, 2010). As a result, there is nowadays a less anthropocentric view of human and animal abilities. Nonetheless, when we widen the scope from humans to a range of species, some fundamental questions about the nature of intelligence become more difficult: what intelligence really is, exactly what cognitive tests measure, which factors there are and whether the relations among cognitive abilities are only valid for a few species or general for the entire animal kingdom. Despite the enormous progress in the past decades, comparative cognition, psychometrics and related disciplines are still very far from drawing a comprehensive and accurate chart of the set of cognitive abilities in the animal kingdom (including humans) and the degree to which each is present in each species (and in each stage of development).

If this is already a vast sea for exploration, there is an uncharted ocean of cosmic dimension: the set of all possible *machines*, with any conceivable (computable) behaviour. We call this set the *machine kingdom*. Any computable behaviour is possible —at least in theory—, including the emulation of any animal (an *animat*, Wilson, 1991; Webb, 2009). While the diversity and richness of behaviour is still much higher in natural systems than state-of-the-art artificial systems, machines can be designed (or conceived of) without any constraint related to evolution niches and species survival. Actually, computer systems are becoming more and more complex and unpredictable (Bullock & Cliff, 2004). This is not different for robots and other cognitive systems in artificial intelligence (AI), as some properties may emerge as the result of a learning process in an environment or as the outcome of some kind of artificial evolution. This suggests that there is an increasing need of tools for evaluating the behaviour (e.g., the abilities) of all these artefacts.

Given this need, do we have a well-founded and functional set of cognitive tests for bots, robots, artificial agents, avatars, 'animats', etc.? There are certainly many anthropocentric approaches, such as the the Turing Test and variants (Turing, 1950; Oppy & Dowe, 2011), the more recent CAPTCHAs (von Ahn et al., 2004) and the mere use of human IQ tests for machines (Sanghi & Dowe, 2003; Detterman, 2011; Dowe & Hernández-Orallo, 2012). We will review some of these works in subsequent sections and discuss their limitations. There is also an incipient line of research where algorithmic information theory plays a crucial role, as we will see. Despite this more mathematical approach, the evaluation of *artificial* cognitive systems is still much more immature than the evaluation of *natural* cognitive systems in psychometrics, comparative psychology and ethology.

In this paper we claim that the statement of a more general problem, the measurement of cognitive abilities for any kind of individual or collective —either artificial, biological or hybrid— may be advantageous over the particular problems dealt with separately by human psychometrics, animal evaluation and AI metrics. Thus, we propose here the notion of 'universal psychometrics' as a subject of study, and eventually a discipline, that would focus on the measurement of any interactive (cognitive) system. In this spirit, and with the general view of the machine kingdom, this paper re-explores and, in most cases, mathematically formalises the concept of cognitive ability, evaluable subject, task choice, interface, difficulty, factorisation, cognitive test, administration, scales, etc. One of the key concepts is the definition of *cognitive task class*, as a weighting (more formally a distribution or probability

measure) of tasks. This is useful to see what a cognitive ability can and, most especially, cannot be, by using some results about task distributions, the no-free-lunch theorems and the universal distributions.

The paper is organised as follows. Section 2 describes how the area of analysis has been enlarged from human evaluation to animal evaluation, with a natural further generalisation in the realm of machines, known as the machine kingdom. We define what an evaluable subject in the machine kingdom is and we give a definition of universal psychometrics. Section 3 overviews the different approaches taken in human psychometrics, animal evaluation in comparative cognition and machine evaluation. Section 4 focusses on what is being measured, properly defining the notion of cognitive task and interface between a subject and a task. Here we also develop the notion of task class and distribution, provide some examples and deal with the notion of task difficulty. Section 5 discusses how expected performance can be approximated for a task class, discusses some other possible indicators and shows decompositions and curves based on difficulty. Section 6 focusses on the exploration of the machine kingdom, introducing a general adaptive test, the use of populations or reference machines for scaling, and the notion of universal test. Section 7 investigates how the space of abilities can be explored around the concepts of validity and the analysis of relations and hierarchies among abilities. It touches upon the notion of general ability and intelligence, and its connections with the so-called no-free-lunch theorems and related results. Section 8 closes the paper with a discussion about the possible objections to universal psychometrics, its possibilities and applicability, and the main contributions of this work.

2. On evaluable subjects: the Machine Kingdom

One of the most important contributions that science makes in many disciplines is the categorisation and classification of phenomena, from rocks to planets, from subatomic particles to plants, etc. Having a good account, a taxonomy, of what we have is linked to the more difficult question of analysing the laws that govern these things. The first thing to do in any discipline is then to define this "universe" of subjects[1] and try to introduce a taxonomy for them.

If we are dealing with measuring cognitive abilities, we need to define precisely what systems can be measured and what abilities can be measured. Next we deal with the subjects. In section 4, we will deal with the abilities.

2.1. The Machine Kingdom

When dealing with the subjects of study —of evaluation, in this case—, we typically encounter a highly fragmented view of the disciplines and procedures depending on the kind of subject —humans, non-human animals[2] and non-biological machines—, as seen in figure 1.

---

[1]If universal psychometrics aims to be truly universal, any kind of cognitive system could be a possible subject to be evaluated. We will be as broad as possible, including humans, other living beings, machines, either on Earth or elsewhere, hybrids of any of these (such as humans augmented by objects and devices), and of course groups or communities of these, and the inter-connection (or communication) protocols making them work as a group.

[2]In what follows, we use the term 'animal' but we do not exclude, in principle, other living organisms, such as plants or fungi.
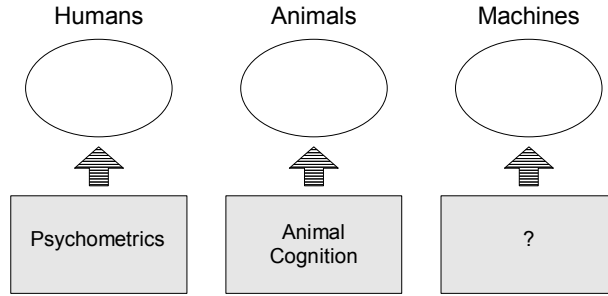
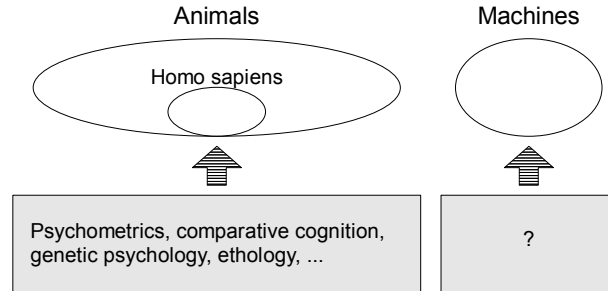Figure 1: Human, animal and machine intelligence are considered separately.

Figure 2: Current view as human intelligence is more and more frequently seen in the context of animal intelligence. Machine intelligence is still considered separately.

The previous picture is simplistic in many ways, since there is an increasing overlap between psychometrics and comparative cognition and ethology, and human and animal cognition in general. The picture has a more accurate representation at present as shown in figure 2. This view is increasingly more popular[3] in comparative cognition (see, e.g., Herrmann et al., 2007, 2010).

Of course, if we see the move from figure 1 to figure 2, it is natural to expect that the worlds of animals and machines could also be integrated in some way. In fact, many theories of cognition are full of computational models, and the metaphor of the computer for humans and non-human animals is already a tradition in many cognitive sciences. However, by integration we mean something more than establishing some parallels. For instance, the area of artificial life and some fields in the area of adaptive behaviour try to reproduce the structures of the left of figures 1 and 2 on their right. The notion of *animat* (Wilson, 1991; Webb, 2009) is the most representative and clear example of this parallel. With this parallel, robotic androids (humanoids) would just be special cases of the concept of animat (see, e.g., Di Paolo, August 2009).

Is this dualistic view with *animats* in a virtual world and animals in the physical world the right picture? We think it is not. The existence of robots in the physical world and people in virtual worlds (Internet, games, etc.) makes this separate scenario unsound and unrealistic. Furthermore, the previous picture is problematic in many ways. What about a human who is assisted with a non-electronic or (possibly) electronic device (or both) —or

---

[3]The issue of *Homo sapiens* being contained within the animal kingdom is one which most —but surely not all— are totally happy with, even though there is overwhelming evidence since Alfred Russel Wallace and Charles R. Darwin were to advance this.
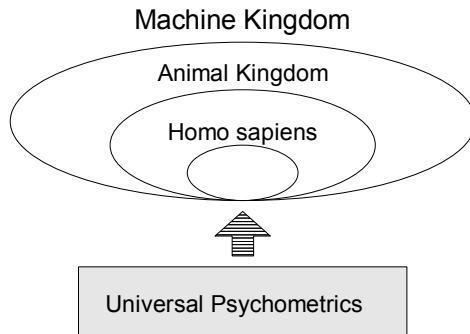
Figure 3: The view of the realm of evaluable agents for universal psychometrics.

a collective of humans, animals and machines? Where do we put them? Which technique should we use to evaluate them?

The key issue about a proper integration is simply the well-known Church-Turing thesis. If this thesis is assumed, everything that is effectively computable is computable by a Turing machine (or any equivalent computing formalism). This thesis is almost universally accepted. The leap comes when we apply this thesis to animals, or, more properly, we apply the *physical* Church-Turing thesis. The physical Church-Turing thesis states that "every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means" (Deutsch, 1985), i.e., a resource-bounded Turing machine[4].

This is compatible with considering animals non-deterministic, since physical machines can be made non-deterministic by any external (physical) source of randomness, such as physically random number generators (Wallace, 1990; Brent, 2008). On some occasions, we explain with non-determinism some effects which can easily be explained with deterministic machines. For instance, for deterministic machines, it is virtually impossible in practice to know their behaviour into the future (because of chaotic effects and also because these machines interact with the real world).

According to all this, the picture simplifies extremely, as shown in figure 3. Biological and robotic systems are in the machine kingdom as well as *animats* and other kinds of virtual agents in virtual environments, hybrids or collectives, provided they are bound to a physical mechanism (electronic, molecular or other).

The best way to see exactly what objects compose the machine kingdom is by introducing a formal definition of interactive system. One of the most general, powerful (and elaborate) models of interactive processes is said to be continuous-time, continuous-state and continuous-action Partially-Observable Markov Decision Processes (MDPs). We can even generalise this by considering any possible domain set for time, states[5], observations and actions.

Definition 1. *An interactive system is defined as a tuple $\langle \mathcal{T}, \mathcal{S}, \mathcal{O}, \mathcal{I}, \dot{s}, \dot{o} \rangle$, where $\mathcal{T}$ is the time space, $\mathcal{S}$ is the state space, $\mathcal{O}$ is the output space, $\mathcal{I}$ is the input space, $\dot{s}(s, i)$ is a*

---

[4]This leap is not universally accepted, with many arguable reasons against (chaos theory, quantum computations, incomputability of the mind, etc.).

[5]Using an appropriately (infinite) state domain, any process which does not follow the Markov property can still be modelled, since the state can be used as memory and $\Delta \mathcal{S}$ as partial updates to that memory.

transition rate function: $\mathcal{S} \times \mathcal{I} \to \Delta\mathcal{S}$, and $\dot{o}(s,i)$ is an output function: $\mathcal{S} \times \mathcal{I} \to \mathcal{O}$.

An interactive system can be non-deterministic (e.g., functions $\dot{s}$ and $\dot{o}$ can be non-deterministic). Interactive systems are usually referred to as 'agents', and outputs are called actions, and inputs are called perceptions or observations.

As we have discussed above, we will only work with (physically) computable agents. It is easy to define appropriate sets for time, states, outputs and inputs such that agents are computable. More precisely, we can use recursively enumerable sets such as $\Sigma = \mathbb{B}^*$, where $\mathbb{B} = \{0,1\}$, i.e. Σ is the set of (possibly empty) binary strings, and the set of rational numbers (denoted by $\mathbb{Q}$). From here, we can use a discrete time[6] in $\mathcal{T} = \mathbb{Q}$ (instead of the set of real numbers, $\mathbb{R}$). The other spaces could be defined similarly: $\mathcal{S} = \Sigma$, $\mathcal{O} = \Sigma$ and $\mathcal{I} = \Sigma$, or restricted to finite sets.

Definition 2. *The machine kingdom (denoted by Ω), is the set of all physically[7] computable interactive systems (or agents).*

The notion of machine kingdom makes some recent works in comparative cognition, adaptive behaviour and robotics more meaningful. For instance, in (Barrera et al., 2011), rats and a specific kind of robot are compared on the same tasks, as if they were two different species. These kinds of comparisons between animals and *animats* will be more and more common in the near future.

The integrated view helps us realise that even though we already knew that the diversity of the animal kingdom is enormous, the diversity of the machine kingdom appears to be unlimited.

2.2. Definition of Universal Psychometrics

The character, size and diversity of the set defined by definition 2 prompts a new field of study, known as universal psychometrics, which is defined as follows:

Definition 3. *Universal Psychometrics is the analysis and development of measurement techniques and tools for the evaluation of* cognitive abilities[8] *of subjects in the* machine kingdom.

---

[6]For the physical world, it might make more sense to (e.g.) have multiples of Planck time using $\mathbb{N}$. However, since seconds are more customary, we will use $\mathbb{Q}$.

[7]Elements in the machine kingdom are not abstract things, but physically-realisable things. The condition of being physically computable is formally stricter than being computable, since it assumes some kind of resource bounds on computation, as the physical Church-Turing thesis does. Similarly, it also entails a stricter definition than being recursively enumerable, namely being computably enumerable with physical resources. A similar appreciation can be made about all the domain sets which compose an interactive system. This means that the set of all agents, i.e., the *machine kingdom* Ω, is a recursively enumerable set with physical resources. We could define the machine kingdom in a more general way, by considering any conceivable machine, physically realisable or not, or several models of 'hypercomputation' (Syropoulos, 2008). We could also do this with the notion of cognitive ability, which will be seen in section 4. Technically, this would lead to a more general (or completely universal) notion of universal psychometrics, and could also dissipate some concerns about the animal kingdom being included in the machine kingdom (as defined here), possibly related to the famous Lucas/Penrose argument. This eventual generalisation is appealing, but we prefer to consider a more restrictive (but already huge) setting, at least for the rest of this paper.

[8]Other traits related to personality or emotions are also possible. We exclude them here for simplicity.

The definition depends on a proper definition of cognitive ability that we will see in section 4, but we can understand cognitive abilities constructed over sets of measurable information-processing tasks. Abilities such as playing basketball well or having a good pronunciation are not purely information-processing processes, since they rely on physical abilities. On the contrary, playing chess well and being able to do written-text translation are (almost) purely cognitive abilities.

Note that the traditional notion of subject in the animal kingdom considers that an individual is usually seen as a single entity throughout its life, even though its cognitive development may give different measurements at different periods in life (e.g., from childhood to adulthood, or after an accident or illness). In the generalised case of the 'machine kingdom' this is more complex since, e.g., we can migrate an AI system to a faster machine, with some of its abilities immediately being improved, while still being the same 'algorithm'. As for definition 2, a subject is a program in a physical machine, so any (small) change to the program or to the machine (including, e.g., learning) makes it —technically— a different subject[9]. The definition implies that abilities related to speed can be perfectly measured (since we have not used the notion of asynchronous interactive discrete-time systems, but a more general setting in definition 2). And, of course, we will still be able to define tasks and abilities for which time is not taken into account.

Another issue related to measurement is the generality of definition 2. We cannot expect that subjects in Ω must necessarily be "agents with goals" or that there must be a way to condition their actions. In some cases, this will be possible, but in some others we will not discover a way to do that, or it may simply be impossible. This highlights the relevance of the *interface* between a subject and a task, which is more important the more diverse the population we want to analyse is. We will introduce proper definitions of interface and task in subsequent sections.

There are many (more) issues when trying to evaluate subjects in the machine kingdom. We will deal with many of them in this paper. Before getting into more details, the very (simple) idea of universal psychometrics helps to show that some techniques concerning measuring cognitive abilities have been derived in a very specific way (e.g., human psychometrics) but can be used in a more general way if properly generalised (e.g., computerised adaptive testing, use of rewards, etc.). It would be good to bring all the knowledge, techniques and experience of psychometrics and comparative cognition for the corresponding measurement of the (broader) machine kingdom. Similarly, there is also an important flow that could be taken from the general approach recently blossoming in machine evaluation to human and non-human animal evaluation, such as the derivation of task difficulty, the use of formal definitions and tests of intelligence, etc. All these techniques from different fields are reviewed in the following section.

---

[9]This has to be much more explicit for machines in general than for animals, since for animals —and humans in particular— their cognitive abilities are assumed to be relatively constant in two different (but reasonably close) periods of time, especially when we refer to fluid —versus crystallised— abilities. For machines we do not know how much their program can change in a very short period of time, so we have to be more careful when assuming that fluid abilities for the same machine remain constant at different times (or even during an experiment). This is related to the notion of potential intelligence we will discuss in section 6.4.

3. Background

Once we have clarified the realm of subjects we want to measure, we need to analyse the theories, techniques and methodologies which have been used in each parcel, with a special emphasis on those which might be still applicable and more useful for universal psychometrics.

3.1. Evaluating humans

Humans have been evaluated by other humans in all periods of history, but it is not until the 20th century that psychometrics was established as a scientific discipline. Psychometrics not only measures abilities (cognitive skills), but also traits (personality) and attitudes (social views and opinions). In this paper we will focus on the measurement of cognitive abilities.

Regarding cognitive abilities, psychometrics developed the crucial distinction between "culture-fair" versus "culture-bound" tests, and the notions of "fluid abilities" versus "crystallised abilities" (see, e.g., Sternberg, 2000; Eysenck, 2007). Thorndike's theory (1927) introduced a distinction between the *width* which is attained on a certain level of difficulty (e.g., the percentage of tasks at this level which are correctly solved) and the *altitude* (the highest difficulty a subject can reach, setting a minimum width threshold). From these two concepts, an aggregate performance result was obtained, called "area" by Thorndike. Although these terms are not so common nowadays, similar ideas still prevail.

Among all the tests developed in psychometrics, intelligence tests, and, most especially, IQ tests have become the most popular and useful, but also the most controversial. Some of the objections are about the very nature of IQ tests. For instance, many philosophers (and some psychometricians) have raised the "subjectivity objection" of the items in IQ tests. This objection says that there is no intrinsic or mathematical reason to choose some of the "correct" answers. Why is 9 the follower of 1, 3, 5, 7? Why not 1 again? Many IQ tests series, progressive matrices, etc., require a "correct" answer. Even though this is related to the inductive inference problem (Solomonoff, 1996; Hernández-Orallo, 2000c)(Dowe, 2008, footnote 128), this objection has not been properly solved in psychometrics. One of the answers which is given is that if the "correct" answer is "explained", most humans agree that this is reasonable. However, by mere human agreement we cannot explain why some other, alternative, explanations can be ruled out.

The set of abilities measured by psychometrics and social sciences is endless. A different question is whether abilities or traits are independent or not. Note that the correlations refer to the species (or any subpopulation, according to age, sex, race, country, etc). When we focus on cognitive abilities, and, more specifically, on the intellect, the relations between abilities are more difficult to analyse. In the case of intelligence, for instance, there is no consensus about how many factors there are and how they should be arranged. This also depends on the theory or structure of (human) intellect being advocated. This is different for each different theory: Spearman-Thurstone, Eysenck, Piaget, Jensen, ... Despite the differences, the most prominent *distinctive* factors are usually a small set, such as inductive reasoning, deductive reasoning, verbal ability, numerical ability, spatial ability, perceptual speed, rote memory and perceptual organisation (Eysenck, 2007). Some other 'factors' or abilities have been appearing and have been a source of long debates, such as the "inspection time" (Sternberg, 2000, chap. 13), which is a very simple information processing ability,

which strongly correlates with some measures of intelligence, in humans. This raises the question of what correlation between abilities really means in humans (or other species), since "inspection time" is clearly different from intelligence, as eyes and hair are different but their colour is related.

Decades of research have not led to convergence of theoretical views. Measuring and understanding intelligence in psychometrics have been governed by factor analysis approaches during decades. Unidimensional and multiple factor theories of intelligence derived from its application, as early as Spearman's two-factor theory of intelligence (Spearman, 1904) — where performance in intelligence tests is the results of the operation of a general factor (g) and specific factors (s) due to unique abilities related to particular tests. Many multiple factor theories followed, such as the one proposed by Kelly (1928), Thurstone's theory of primary mental abilities (Thurstone, 1938), Guilford's three-facet model of the structure of intellect (Guilford, 1967), hierarchical models such as Vernon's and Burt's (Vernon, 1950; Burt, 1949), the hierarchical structure proposed as well by Cattell and Horn (see, e.g., Horn & Cattell, 1967) —with g divided into fluid intelligence (gf) and crystallised intelligence (gc)— and the most recent three-stratum theory of intelligence proposed by Carroll (1993) —with a general factor in Stratum 3. Most IQ tests widely used derived from these models/theories, derived in turn from factor analysis approaches. Some of them are most directly related to a general factor model of intelligence, such as Raven's Progressive Matrices Tests (Raven et al., 1992), and Wechsler scales (e.g., WAIS-R, Wechsler, 1981), while others are associated with the multiple factor theories, such as Thurstone's Primary Mental Abilities Test and the updated Schaie-Thurstone Adult Mental Abilities Test (Schaie, 1985).

Distinct theories have also been proposed for psychological measurement that differ in the way they conceptualise the measurement process. Two of the most important measurement models are Classical Test Theory and Item Response Theory (CTT and IRT respectively henceforth). CTT is still the most influential and widely used in empirical application (chiefly in psychology). The central point of CTT is the concept of "true score" that is conceptualised as the expected value of the observed score. In this respect, Boring's "intelligence is what intelligence tests measure" applies perfectly to this model since, for CTT, someone's intelligence is her expected score on an IQ-test. Some have criticised this approach (Borsboom, 2005), arguing that the underlying philosophy is an ambiguous operationalism in which the meaning of theoretical attributes and the operations and instruments used to measure them are the same thing.

The more recent IRT approach (e.g., Rasch, 1960) is challenging in this respect and supposes an advance in the field. IRT, or latent variable model, is more popular in the academic arena and psychometric research. Based on a realistic philosophy, traits or theoretical attributes are treated as latent variables and are conceptualised as the common cause of observed variables. From this approach, for example, intelligence is seen as the common cause of the responses to a particular IQ test (or intelligence item). An observed score on an IQ test is conceived as a particular measure, index of the latent variable, intelligence (as far as the statistical model, based on the theory of the latent variable, fits the data). One popular IRT model is the *three parameter model* (see, e.g., Lord et al., 1968):

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

which, based on a logistic function, estimates the probability of guessing the item right according to a level θ of proficiency or ability, where parameter a is the discriminative power of the item, b is its difficulty, and c is the binomial floor on the probability of getting an item correct. Parameters a and b are usually inferred experimentally, using a (sub)population as a reference. This is important, since IRT is frequently referred to as a "formal" and "well-founded" way of choosing test items. This does not mean, however, that items and their difficulty are obtained from theoretical principles, but rather that a well-founded theory is used to empirically estimate their parameters. A few recent models following the IRT approach can include parameters to describe the cognitive demand of the items (such as the 'Confirmatory IRT Models' in (Sternberg, 2000, chap. 19)). This has important theoretical implications since the model provides (at least partial) information about what is measured by the items, being therefore better interfaced with cognitive psychological approaches to understanding intelligence.

Computerised Adaptive Testing (CAT) (Wainer, 2000; Weiss, 2011) usually relies on IRT to develop adaptive tests which select the most informative items according to the results on previous items. Adaptive tests are usually more efficient than classical ones, especially when the ability range of the examinee is unknown in advance. However, CAT tests require an accurate estimation of IRT parameters.

Another important concept is the distinction between *reliability* and *validity*. Basically, reliability refers to the precision of a measure, that is, the degree to which a measure is error-free (and whether the ability is consistently measured for the same individual in different moments, e.g., test-retest estimation). In estimating reliability, different measurement models use different approaches (Sternberg, 2000; Rasch, 1960; Embretson & Reise, 2000, chap. 19-21). In CTT, we have different procedures (e.g., test-retest, parallel forms, internal consistency, etc.) and each one of them tells us about different sources of error (e.g., temporal stability, degree of parallelism, sampling error, ...). In IRT, "information functions" are also used to describe the degree of precision of a test at different levels of the estimated ability (or trait). On the other hand, validity is not easy to define. While the concept has evolved significantly (see, e.g., AERA, APA & NCME, 1999; Borsboom et al., 2004), for the rest of the paper we will stick to the traditional (and still popular) concept: "a test is valid if it measures what it is supposed to be measuring" (Kelley, 1927).

Summing up, there are many caveats and objections about psychometrics, but it is widely recognised that psychometrics is able to measure human intelligence and cognitive abilities accurately. The widespread use of psychometrics in education, psychology, psychiatry and human resource selection is still one of the best corroborations of the success of psychometrics.

3.2. Evaluating non-human animals

In a sense, the foundation of animal cognition research has it roots in Charles Darwin's and Alfred Russel Wallace's ideas. Their still controversial hypothesis that humans' and other animals' minds differ just in degree and not in kind still remains as a hot topic in the field. Since the beginning, comparative psychology has focussed in cognitive differences between species. The first studies (e.g., Harlow et al., 1950; Bitterman, 1960) positioned within the framework of *scala naturae*, focussed in searching for *universal* laws of learning.

Behaviourism pushed comparative cognition towards general models of cognition (e.g., Skinner, 1938; Watson, 1967) and focussed on a very few species, with the white rat as the

main subject of study (see MacLean et al., 2011), with less emphasis on species comparison (Shettleworth's paper "Where is the comparison in comparative cognition?", 1993). The last three decades have brought an important evolution in the field (Griffin, 1978). The number of studied species increased significantly and the scope shifted toward other animals, with apes, monkeys, and humans as the most popular ones. This trend was accused of "chimpocentrism" (Beck, 1982), since many studies focussed on apes, but today the spectrum has been widened significantly in the animal kingdom: mammals (cetaceans, dogs, elephants, ...), other vertebrates (birds, fish) and even invertebrates (bees, ants, octopuses, ...). All this is shedding light about convergence and divergence phenomena in the evolution of cognition (Shettleworth, 2009, 2010).

Today, the study of cognition comprises the study of the animal mechanisms for acquiring, processing, storing and acting on information available in the environment. Following (Tomasello & Call, 1997), comparative cognitive studies encompass processes (of acquisition, representation and behavioural control) related to two main domains defined in terms of what aspect of the world they are about: physical and social cognition. The study of the cognition of time, space, numbers, causal reasoning and tools use are typically in the physical cognition domain while processes such as social learning, theory of mind and communication are within the social domain.

There is an important difference between the study of cognitive abilities in human psychometrics and animals. Human abilities are usually linked to the success and performance in some (specific) contexts in life (e.g., academic performance and highly-qualified jobs). From the point of view of animal cognition, cognitive abilities and, of course, intelligence are seen from an evolutionary point of view in terms of environments and ecosystems. For instance, if a species has more long-term memory or a shorter response time, its individuals may be more fitted to capture prey and to identify and escape from predators. All this suggests a complex relation between survival, reproduction, fitness and some cognitive abilities.

For universal psychometrics, the part of animal cognition which is most relevant is the one which comprises the techniques and methodologies which have been devised for evaluating animals. Animal testing is much more difficult than (adult) human testing. The reasons are clear: it is more difficult to give instructions to animals or to coerce them to do something, to keep their attention, etc. This also happens with human testing as well, most especially with children and people with disabilities. The answer for this is typically the use of rewarding protocols, which in the case of animals means some kind of delicatessen.

On many occasions, we have seen that the interface is crucial. If we use the wrong food as a reward for an animal, its performance will be underestimated. This raises a very important point in measurement (and also in universal psychometrics): given a task, the result on a task may be underestimated with a bad interface design. It is frequently argued that the poor scores of some animals (or even people) at some tasks are explained by a bad interface design. The reasons are not always easy to spot, since the subject may not like the rewards as much as the tester thought, the subject may want to mock or dupe the tester or feel suspicious about her, the subject may want to use the test to *communicate* something with the tester, or even, the subject may want to *test* the tester (Solomonoff, 1967, sec. 6, The Problem of the Ambitious Subordinate). A bad interface design might also cause the test to overestimate, because it may give more information than needed, or it may give the information in an unfairly advantageous way. On other occasions, the task which is being

evaluated might require another ability which is taken for granted, but is only present in one species but not in others.

Although difficult, evaluating different species in a fair way is still at least largely possible, if done carefully. An example of how different species can be evaluated, one of them being human children, while aiming at discovering the effect of social abilities, can be found here (Herrmann et al., 2007). In this work, the exercises are the same but the interfaces are different, including different rewards for each kind of subject.

A final issue for which animal testing is pioneering as well is in the measurement of collectives. There are studies evaluating ant colonies, bee swarms, and other animal flocks (see, e.g., Strassmann et al., 1997). In these cases the subject is the group, and not the individual. This obliges us to find suitable interfaces, as well as appropriate rewarding and scoring mechanisms, since some individuals might even die during the test as a useful sacrifice for the common 'goal'. This is related as well to the measurement of social abilities.

3.3. Evaluation in Artificial Intelligence

Both psychometrics and comparative psychology are based on the notion of 'species'. While psychometrics evaluates individuals by placing them in normalised scales which are derived from the human population, comparative psychology typically extrapolates the cognitive abilities of a species from some testing on a sample of individuals of that species. In other words, psychometrics focusses on the evaluation of individual differences while comparative cognition generally focusses on the evaluation of differences between species. There are, however, some exceptions, such as (Herrmann et al., 2010), which evaluates individual differences in more than one species.

Things are different for machines. We can say that two machines are identical, or talk about a "series" of machines with similar characteristics (according to their design), but certainly no notion of species which compare or normalise. An interactive machine can be any computable process, for which even the notion of 'individual' is different.

When talking about tests for machines, we need to start with the Turing Test (originally, the imitation game, Turing, 1950) and its many variations (Oppy & Dowe, 2011). In the previous section, we have seen how comparative cognition works. The Turing test tries to put systems in equivalence classes by comparing two different kinds of individuals. In fact, it compares an individual with a species (with one human as a sample). In a way, it is a very special (and rudimentary) kind of comparative test, which could be a very nice experiment in comparative psychology. For instance, how much time does a committee of judges need to tell a bonobo from a chimpanzee (using a teletype communication)? While this might be insightful as an imaginary experiment, nobody is currently doing these kinds of things in practical animal cognition. We must say otherwise in AI, since some people still do think that the Turing Test (or some of its variants) is a good test or even a practical test. It is primarily just a test of humanity, with human judges, but fairly useless for determining and measuring different cognitive abilities.

A related idea for comparing (or telling apart) two different kinds of systems is the notion of CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart, von Ahn et al., 2004), which poses extremely discriminative exercises that are easy for humans but not achievable with current computer technology. These kinds of tests are very useful in real applications but they do not really evaluate cognitive abilities in general.

CAPTCHAs are also 'cracked' regularly, and they need to be replaced by more sophisticated ones, which are currently becoming difficult for humans, thus losing discriminative power.

Apart from these comparative approaches, there are some (although not too many Long & Fox, 2003; Genesereth et al., 2005; Whiteson et al., 2010) specific tests and competitions which have been proposed (and some still running) in the field of AI. All of them, however, are just constructed as a set of specific tasks (learning, planning, robotics, games, ...) and not meant to evaluate general individuals (only individuals conceived for that task). Many of these competitions modify and extend the set of tasks each year in order to cover a broader view of the field and avoid competition-specialisation. However, the 'difficulty' of each task is always quantified or estimated in an informal or ad-hoc way, so it is very difficult to compare results across different algorithms and competitions.

An exceptional area in artificial intelligence where evaluation practices are closer to the kind of evaluation performed in comparative cognition is Reinforcement Learning (RL) (Sutton & Barto, 1998; Woergoetter & Porr, 2008). RL is a proper and general setting for defining and analysing learning agents which interact with an environment through the use of observations, actions and rewards. Hence, RL is not strictly limited to AI agents; non-human animals and humans can be understood in this setting, most especially in the context of evaluation. In fact, the theory of reinforcement learning is now overlapping with areas dealing with adaptive behaviour and cognition in general. One important (and, at the same time, complex) idea in RL is how rewards are aggregated, as well as the problem of exploration-exploitation, which is usually solved by the use of discounted reward utility functions or other aggregated functions of performance (Hernández-Orallo, 2010). The tests in these areas are generally focussed on specific tasks, as in the rest of AI. However, tasks are not predefined but also learned through rewards, as happens with animal testing. Nonetheless, the goal in animal testing is not always to make the animal learn to do something, but to discover the natural cognitive way to proceed.

A straightforward but very different idea for evaluating machines is the use of IQ tests (or, more generally, human intelligence tests and tests of other cognitive abilities). This has been indirectly suggested many times in AI, especially in its beginning. For instance, Evans (1964) cracked some exercises similar to IQ tests using his program *ANALOGY*. The goal of this line of research was set by Newell (1973): "to construct a single program that would take a standard intelligence test". Recently, this approach has been "resurrected" (Bringsjord, 2011) under the name "Psychometric AI" (Bringsjord & Schimanski, 2003; Bringsjord, 2011), which first states that"some agent is intelligent if and only if it excels at *all*[10] established, validated tests of intelligence" (Bringsjord, 2011) and, second, defines "psychometric AI [as] the field devoted to building [these agents]" (Bringsjord, 2011). Remarkably, this is *not* "AI psychometrics", which would be the field devoted to *evaluating* these agents. While "psychometric AI" may commendably revive the relation between psychometrics and AI, it reproduces the old circularity in psychometrics: "intelligence is what the intelligence tests measure". Ultimately, the selection of "all established, validated tests of intelligence and mental ability, [...], tests of artistic and literary creativity, mechanical ability, and so on" (Bringsjord, 2011) relies on an agreement of what an 'established, validated test' is and a

---

[10]Emphasis is ours.

selection among "the myriad tests that psychometricians have validated" (Bringsjord & Schimanski, 2003). A similar route is traced by Douglas K. Detterman, editor of the *Intelligence* journal, who recently announced (Detterman, 2011) a challenge (originally to Watson, the AI winner of the *Jeopardy!* TV quiz show), suggesting the use of IQ tests to evaluate the whole field of artificial intelligence. Since a program passing a *single* psychometric test does not mean that the program is intelligent (emphatically shown in (Sanghi & Dowe, 2003)), the goal seems then to be redefined as passing a whole *battery* of tests. This (and psychometric AI) conceals the fact that a set or battery of tests is just *a* larger (possibly more heterogeneous) test.

We think that the perspective of universal psychometrics we shape in this paper gives some clues as to why the above-mentioned paths are wrong. IQ tests (and many other cognitive tests) are not for machines, at least not yet, as has been further elaborated in (Dowe & Hernández-Orallo, 2012). In fact, these tests, and many other psychometric tests for other cognitive abilities are not (yet) even useful for animals. They can be used as an ultimate goal, as an inspiration or even as a measure of progress in AI, but they cannot be used, in the long term, as valid tests for universal psychometrics. The reason is that IQ tests are —at least currently— anthropocentric. One problem is that they assume many things about the evaluated subject. A second problem is that we do not know exactly what these tests measure, since —in general— there is no formal theory or model behind how the exercises are generated. Third, there is no general mathematical account of the difficulty of exercises, apart from a normalisation to the human population. Fourth, it has been shown that IQ tests can be fooled by machines[11]. All this does not mean that the (human) psychometric approach is not useful to develop tests for machines, but that some principles must be changed if the remaining available tools are to be preserved.

A different *and principled* approach started in the late 1990s, while its origins can be traced back to Chaitin's invitation to "develop formal definitions of intelligence and measures of its various components [using algorithmic information theory]" (Chaitin, 1982, sec. 6, item f). An *induction-enhanced* Turing Test was proposed (Dowe & Hajek, 1997, 1998) in such a way that a general inductive ability can be indirectly measured by the test. The importance was not that any kind of ability could be included in the Turing Test, but that this ability could be formalised. This was done by defining inductive inference à la MML (Wallace & Boulton, 1968). The Minimum Message Length (MML) principle advocates for (two-part) compression as a way to perform inductive inference (i.e., learning) and, ultimately, intelligence (Wallace & Dowe, 1999; Wallace, 2005; Dowe, 2011, sec. 7.3). MML is closely related to Solomonoff's theory of prediction and Kolmogorov complexity (Wallace & Dowe, 1999; Li & Vitányi, 2008).

Independently, a new intelligence test (*C*-test) (Hernández-Orallo & Minaya-Collado, 1998; Hernández-Orallo, 2000a) was derived as sequence prediction problems which were gen-

---

[11]In particular, in 2003 a Perl computer program —with just 960 lines of code— performed quite well on standard human IQ tests (Sanghi & Dowe, 2003), even surpassing the average score (of 100) on some tests. The purpose of the experiment was not to show that this program was intelligent —or to make progress in AI, as the psychometric AI approach vindicates— but to show that IQ tests were not for machines (Dowe & Hernández-Orallo, 2012). In fact, for every single new IQ test which is issued, a machine can be devised to pass the "kind" of tasks which comprise the test, something which is also happening with CAPTCHAs.

erated by a universal distribution (Solomonoff, 1964) and the related algorithmic information theory (a.k.a. Kolmogorov complexity (Li & Vitányi, 2008) or Solomonoff-Kolmogorov complexity). The C-tests estimated the intelligence of a subject (or, perhaps, more precisely, its inductive inference ability) by using sequences whose complexity was mathematically derived from constructs derived from Kolmogorov complexity. These exercises were very similar to those found in some IQ tests, but here they were created from computational principles. In fact, the test was administered to humans, and the results strongly correlated with some IQ tests (Hernández-Orallo & Minaya-Collado, 1998). For the authors, the test was taken as a definition of intelligence. As a result, 'intelligence' was no longer (only) what the IQ tests measure (Boring, 1923), but (also) a precise mathematical concept. In addition, this work "solved" the traditional subjectivity objection of the items in IQ tests, i.e., since the continuation of each sequence was derived from its shortest explanation, although this is still slightly subjective relative to the reference machine used. However, this test only measured one cognitive ability and its presentation was too narrow to be a general test. Consequently, these ideas were extended to other cognitive abilities and other kinds of subjects in (Hernández-Orallo, 2000a) by the introduction of other 'factors', and the suggestion of using interactive tasks where "rewards and penalties could be used instead", as in reinforcement learning.

Similar ideas followed relating compression and intelligence. Compression tests were proposed as a test for artificial intelligence (Mahoney, 1999), arguing that "optimal text compression is a harder problem than artificial intelligence as defined by Turing's". Nonetheless, the fact that there is a connection between compression and intelligence does not mean that intelligence can be just defined as compression ability (see, e.g. Dowe et al., 2011 for a full discussion on this).

(Dobrev, 2000, 2005) proposed a definition of (artificial) intelligence as an aggregate of performance in a wide range of worlds, where the set of worlds is described by Turing machines and bounded by Kolmogorov complexity. Also, the agent with smallest Kolmogorov complexity is said to be the best agent in this set of worlds. This agent, called "AI", is proposed to be found by a Levin search. Another approach following this trace is a refurbished notion of "universal intelligence", formalised by (Legg & Hutter, 2007), also using Kolmogorov complexity, Solomonoff's universal distributions, and an aggregate of performances in all the possible environments. However, several issues arise from here, such as (e.g., see Hernández-Orallo & Dowe, 2010, secs. 3.3 and 4) the apparent neglect of time, the use of several sources of infiniteness and other counterintuitive consequences when trying to derive a test from these ideas.

In (Hernández-Orallo & Dowe, 2010), the problem was focussed back again to finding good *tests* of machine intelligence, and to the notion of "universal intelligence test", where "universal" stands for any kind of subject. Some of the problems in previous approaches are spotted, and solutions proposed. Also, this paper extends the notion of machine intelligence test to adaptive scenarios, including time and adjusting the complexity of items to the subject (instead of using a universal distribution for this). The notion of adaptive, anytime intelligence test, was introduced, with clear similarities to Computerised Adaptive Testing in psychometrics. Other notions such as the need for items to be discriminative, and a scale of results where random agents are placed at 0, are firstly introduced in this work. This paper also suggests that this setting could be used for any realm in AI, e.g. games, machine learning, etc., which is a first step to reconcile the extreme fragmentation of AI subfields

with a unified measuring setting. The setting has been used to develop preliminary tests for both humans and machines (see, e.g., Insa-Cabrera et al., 2011; Hernández-Orallo et al., 2011), whose results show that there still remains a *very* long way ahead[12].

4. On cognitive abilities

After the previous overview of related disciplines and ideas, we are ready to introduce one of the main components of the definition of universal psychometrics: the notion of cognitive ability.

Definition 4. *A cognitive ability is a property of individuals in the machine kingdom Ω which allows them to perform well in a* selection *of information-processing tasks.*

This definition implies that abilities are constructs while tasks are instruments. Cognitive abilities can be inferred by the performance on tasks —among other instruments— , but they are not tasks themselves. In fact, it is very difficult (but arguably not impossible) to find a set of tasks which corresponds uniquely with an ability, because a task usually involves several abilities. We will discuss this later in the paper. For the moment, let us start with the formalisation of task.

4.1. Cognitive task

An interesting duality of interaction is that tasks can also be seen as interactive processes (see definition 1). In this case, inputs are known as actions and outputs are known as observations. There are many formalisms to define this kind of interactive processes, with, again, (Partially-Observable) Markov Decision Processes (MDP) being very popular and general. In many cases, discrete time, alternating interaction[13], and finite states are assumed. In this paper we are interested in the most general kind of processes. For instance, many perception abilities or control abilities cannot be (properly) modelled with a discrete, alternating time. By 'most general', we mean environments such as those we are used to seeing in real cognitive tasks, virtual problems or games (such as video games), where very general state, action and observation domains can be used, and time is asynchronous (agents can perform actions at any time). Such a general view is precisely definition 1, which can be used to define the notion of cognitive task as follows:

Definition 5. *A cognitive task is defined as a physically[14] computable interactive system $\langle \mathcal{T}, \mathcal{S}, \mathcal{O}, \mathcal{I}, \dot{s}, \dot{o} \rangle$ following definition 1 (with outputs being observations and inputs being ac-*

---

[12]A test may work well for a given population or kind of subject, e.g., humans, such as psychometrics tests or the C-tests (Hernández-Orallo, 2000a), or, a family of reinforcement learning agents, such as (Insa-Cabrera et al., 2011; Legg & Veness, 2011). However, the challenge of a universal test is really on its application to different kinds of subjects (e.g., humans and machines) and observe whether the results represent a coherent scale.

[13]By alternating interaction we mean that agent and environment alternate on their inputs and outputs, as usual in discrete-time reinforcement learning.

[14]Again, we set the constraint of being physically constructible (computable), while it is true that some problems, such as the decidability of the truth of any input first-order formula, could be a 'task'. However, these problems are undecidable and not constructible as a set of exercises (since we cannot know the answer for all of them). Again, we do not discard that the notion of cognitive task could be further generalised, but we stick to this simpler version for the rest of the paper.

tions), and a score transition function $\dot{u} : \mathcal{U} \times \mathcal{S} \times \mathcal{O} \rightarrow \Delta\mathcal{U}$, *with* $\mathcal{U}$ *being a bounded set in* $\mathbb{Q}$.

A cognitive task is then an interactive system with a *score* function $\dot{u}$ over a domain of scores $\mathcal{U}$. Tasks can be non-deterministic (e.g. functions $\dot{s}, \dot{o}, \dot{u}$ can be non-deterministic). As we did for agents, the set of all tasks (denoted by $\mathcal{M}$) is also assumed to be recursively enumerable.

Given the previous definition, a couple of examples below will show how it can be used to specify very different cognitive tasks. The first task we use for illustration is number comparison. The goal is to compare two numbers (shown as panels with some of their lights on) and select the greatest one, which is a common test for children and some animals (see, e.g., Shettleworth, 2010). This is a one-exercise task, with no time-limit to answer:

Example 1. *Consider the task 'number comparison'* μ_{nc} *defined with the following domains: time* $\mathcal{T} = \mathbb{Q}$, *outputs* $\mathcal{O} = \langle \mathbb{B}^m, \mathbb{B}^n \rangle$, *inputs* $\mathcal{I} = \{\textbf{left}, \textbf{right}, \textbf{none}\}$, *scores* $\mathcal{U} = \{-1, 0, 1\}$, *and states* $\mathcal{S} = \langle \{\textbf{answered}, \textbf{waiting}\}, \mathbb{B}^m, \mathbb{B}^n, \{-1, 0, 1\} \rangle$ *with the second and the third terms in the tuple storing the numbers and the fourth term storing the score. The states*[15] *evolve according to:*

- $\dot{s}(\langle \textbf{waiting}, x, y, e \rangle, \textbf{none}) = \langle \textbf{waiting}, x, y, e \rangle$,

- $\dot{s}(\langle \textbf{waiting}, x, y, e \rangle, \textbf{left}) = \langle \textbf{answered}, x, y, e' \rangle$, *with* $x = \langle x^{(1)}, x^{(2)}, ..., x^{(m)} \rangle$ *and* $y = \langle y^{(1)}, y^{(2)}, ..., y^{(n)} \rangle$, $e' = 1$ *if* $\sum_{i=1}^{m} x^{(i)} > \sum_{i=1}^{n} y^{(i)}$, $e' = -1$ *if* $\sum_{i=1}^{m} x^{(i)} < \sum_{i=1}^{n} y^{(i)}$ *and* $e' = 0$ *otherwise.*

- $\dot{s}(\langle \textbf{waiting}, x, y, e \rangle, \textbf{right}) = \langle \textbf{answered}, x, y, e' \rangle$, *with* $x = \langle x^{(1)}, x^{(2)}, ..., x^{(m)} \rangle$ *and* $y = \langle y^{(1)}, y^{(2)}, ..., y^{(n)} \rangle$, $e' = 1$ *if* $\sum_{i=1}^{m} x^{(i)} < \sum_{i=1}^{n} y^{(i)}$, $e' = -1$ *if* $\sum_{i=1}^{m} x^{(i)} > \sum_{i=1}^{n} y^{(i)}$ *and* $e' = 0$ *otherwise.*

- $\dot{s}(\langle \textbf{answered}, x, y, e \rangle, a) = \langle \textbf{answered}, x, y, e \rangle$ *with* a *being any action.*

where each $x^{(i)}$ *(or* $y^{(i)}$*) represents a light on the 'panel'* x *(respectively,* y*), and can be either 0 (off) or 1 (on)*[16]. *The observations in* $\langle \mathbb{B}^m, \mathbb{B}^n \rangle$, *where each argument represents a panel, evolve according to:*

- $\dot{o}(\langle b, x, y, e \rangle, a) = \langle x, y \rangle$.

Score evolves as follows:

- $\dot{u}(u, \langle \textbf{waiting}, x, y, e \rangle, z) = 0$.

---

[15]States are quadruples here. For simplicity, the transition functions $\dot{s}$ and $\dot{u}$ are not expressed as differential equations or modifications (over $\Delta\mathcal{S}$ or $\Delta\mathcal{U}$), but as regular functions (over $\mathcal{S}$ or $\mathcal{U}$).

[16]This example uses a non-unique tuple-representation for numbers, so, e.g., for $m = n = 5$ we have that $\langle 1, 0, 1, 1, 0 \rangle$ and $\langle 0, 0, 1, 1, 1 \rangle$ represent the same number, 3. This representation is much more culture independent than a binary or decimal representation for numbers. In fact, when evaluating numerical abilities with animals and children, this kind of representation (or a similar one, such as a cup or plate with a number of items) is used.

- $\dot{u}(u, \langle\textbf{answered}, x, y, e\rangle, z) = e$.

The initial state is randomly derived from $\langle\textbf{waiting}, U(\{0,1\})^m, U(\{0,1\})^n, 0\rangle$, which represents that both the left and the right panels switch its m and n lights according to a uniform distribution $U(\{0,1\})$. Score is kept forever after the guess or answer. This task requires us to set particular values for m and n, which also set how big the numbers that must be compared can be.

The second example corresponds to a variation of a popular family of tasks in human intelligence, the measurement of 'reaction time' to some stimulus or signal, which may or may not require some reasoning. Here, we will formalise one of the simplest possibilities. A signal (e.g., a light) is switched on at a random time or frequency. If the subject is able to react quickly, it gets positive *rewards*. If the subject reacts too late (or before the light is on) then rewards are negative. Scores and rewards are different things. Rewards are part of the observations, working as indicators for the subject. In this example scores are calculated as an aggregation of rewards, but this does not need to be so, in general. This example is presented as a never-ending task, i.e., there are infinitely many attempts. In particular, when the light is switched on and the reward is given, another exercise is proposed, and so on indefinitely.

Example 2. *Consider the task 'reaction time' μ_{rt} defined with the following domains: time $\mathcal{T}$ = $\mathbb{Q}$, outputs $\mathcal{O} = \langle\{\textbf{light-on}, \textbf{light-off}\}, \mathbb{Q}\rangle$ with the second term of the observation meant to be understood as a reward, inputs $\mathcal{I} = \{\textbf{press}, \textbf{release}\}$, scores $\mathcal{U} = \mathbb{Q}$, and, again, states are represented as tuples $\mathcal{S} = \langle\{\textbf{rewarding}, \textbf{no-rewarding}\}, \mathbb{Q}, \mathbb{Q}\rangle$, where the second and third terms in the state tuple are used as time counters. The states evolve according to:*

- $\dot{s}(\langle\textbf{no-rewarding}, x, y\rangle, \textbf{release}) = \langle\textbf{no-rewarding}, x - dt, 0\rangle$,

- $\dot{s}(\langle\textbf{no-rewarding}, x, y\rangle, \textbf{press}) = \langle\textbf{rewarding}, x - dt, \tau_r\rangle$,

- $\dot{s}(\langle\textbf{rewarding}, x, y\rangle, a) = \langle\textbf{rewarding}, x, y - dt\rangle$, *if $y > 0$*

- $\dot{s}(\langle\textbf{rewarding}, x, y\rangle, a) = \langle\textbf{no-rewarding}, Exp(\lambda), 0\rangle$, *if $y \leq 0$, with Exp being the exponential distribution.*

The observations evolve according to:

- $\dot{o}(\langle\textbf{no-rewarding}, x, y\rangle, a) = \langle\textbf{light-on}, 0\rangle$ *if $x \leq 0$, and $\langle\textbf{light-off}, 0\rangle$ otherwise.*

- $\dot{o}(\langle\textbf{rewarding}, x, y\rangle, a) = \langle\textbf{light-on}, \frac{1}{\lambda} - x\rangle$ *if $x \leq 0$, and $\langle\textbf{light-off}, -\frac{1}{\lambda}\rangle$ otherwise.*

Score evolves as a function of rewards (here, r, a part of the observations):

- $\dot{u}(u, s, \langle l, r\rangle) = \frac{u(t - dt) + r \cdot dt}{t} = u + \frac{dt}{t}(r - u)$

The initial state is $\langle\textbf{no-rewarding}, Exp(\lambda), 0\rangle$ and the initial score $u = 0$. The values λ and τ_r can be seen as parameters for the task. The first, λ, gauges the average time for signals. For instance, if λ is set to 0.2, it means that the average time for a signal is 5 seconds. The second parameter, τ_r sets the time the reward is active (and the game is temporarily 'stopped'). For instance, it could be set to $1/(10\lambda)$ seconds. This would mean that the reward is enabled for half a second. Note that the reward requires some duration, since the score function can be seen as a normalised integral or sum of rewards so far.
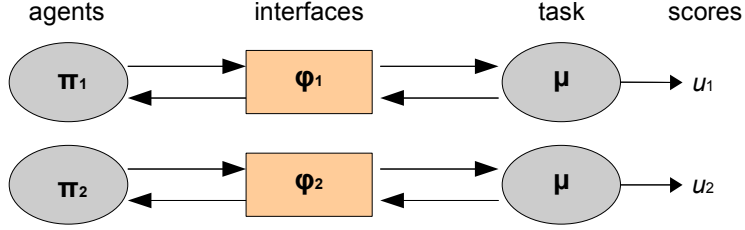
Figure 4: An example with two different agents and interfaces for the same environment.

4.2. Interfaces and episodes

Now that we have proper definitions for agents and tasks, we need to connect them. The way to connect a task with an agent is through an interface:

Definition 6. *Given a task $\mu = \langle \mathcal{T}_\mu, \mathcal{S}_\mu, \mathcal{O}_\mu, \mathcal{I}_\mu, \dot{s}_\mu, \dot{o}_\mu \rangle$ and an agent $\pi = \langle \mathcal{T}_\pi, \mathcal{S}_\pi, \mathcal{O}_\pi, \mathcal{I}_\pi, \dot{s}_\pi, \dot{o}_\pi \rangle$, an interface ϕ between them is a tuple of mappings $\langle \phi_\mathcal{T}, \phi_\mathcal{A}, \phi_\mathcal{Z} \rangle$, with $\phi_\mathcal{T} : \mathcal{T}_\pi \to \mathcal{T}_\mu$ being the time mapping, $\phi_\mathcal{A} : \mathcal{O}_\pi \to \mathcal{I}_\mu$ being the action mapping, $\phi_\mathcal{Z} : \mathcal{I}_\pi \to \mathcal{O}_\mu$ being the observation mapping. Given these mappings, we will take the domains of the task as a reference, just making $\mathcal{T} = \mathcal{T}_\mu$ the time domain, $\mathcal{A} = \mathcal{I}_\mu$ the action domain, and $\mathcal{Z} = \mathcal{O}_\mu$ the observation domain.*

The relation between agents, tasks and interfaces can be seen in figure 4. As we see in definitions 5 and 6, the concept of 'reward' does not exist as a separate entity. In case rewards are used, they are considered part of the task's output and part of the agent's inputs. Note that in reinforcement learning, rewards and observations are considered in a separate way. One straightforward way to work with rewards is to define task's outputs ($\mathcal{Z} = \mathcal{O}_\mu$) as a tuple $\langle \mathcal{L}, \mathcal{R} \rangle$ where the second term is understood as a reward as in example 2. Obviously, in order to use rewards, both task and agent have to treat part of their outputs and inputs (respectively) as a reward.

Note that the score function $\dot{u}$ is not used in the interface. Consequently, the score is not visible to the agent and cannot be calculated from the observation (alone) in general. Any appreciation the task wants to convey to the agent during the interaction must be done using observations (and possibly embedded rewards). In fact, it is important to see that score can be completely independent from rewards. This may sound counterintuitive if compared to typical agents and environments in reinforcement learning, where performance is calculated as an aggregation of rewards. However, it is important to emphasise that reinforcement learning is about 'learning' and many cognitive tasks are not about learning. In addition, on many occasions, even in learning scenarios, the performance cannot be derived as an aggregation of rewards.

It is then important to distinguish between scores and rewards. Scores are the way tasks are evaluated (so a task must always give a score), while rewards (when they are used) are a means to condition or indicate the agent should do some actions and should avoid some other actions. Obviously, the same task can have different results on the same agents if the mapping $\phi_\mathcal{Z}$ (or more properly its inverse, $\phi_\mathcal{Z}^{-1}$) leads to an improper correspondence of rewards to some of the agents' sensory devices or inputs such that this does not have an appropriate motivating effect. The interface is a crucial thing, especially when evaluating a diverse set of agents.

19

Definition 6 does not impose any additional constraint over interfaces being injective, surjective or bijective. However, we think that an appropriate interface should be *bijective*, because only in this case can it give us an isomorphic representation of the problem with no information loss[17]. The following proposition shows this.

Proposition 1. *An interface, as defined in definition 6, cannot introduce additional information about the task and cannot withhold any information if and only if all the mappings are bijective.*

Proof. In order to convey extra information, we would need an input x for which we could choose at least between two possible alternatives y_1 and y_2, such that $\phi_{\mathcal{Z}}^{-1}(x) = y_1$. Since in order to convey information these two alternatives must be possible, we would have $\phi_{\mathcal{Z}}(y_1) = x$ and $\phi_{\mathcal{Z}}(y_2) = x$. This cannot be the case since $\phi_{\mathcal{Z}}$ is a bijection. The other direction of the proof is similar. $\square$

This means that an interface that is injective but not surjective or vice versa (or even neither) will result in a distorted representation of the intended problem —possibly withholding or giving additional information. This does not mean that the interface cannot be helpful for an agent[18].

From here, we can have agents interacting in a task. Since agents and tasks may be non-deterministic, we may have different episodes for the same pair of agent and task.

Definition 7. *An episode is a tuple of functions $\langle s_\mu, s_\pi, a, z, u \rangle$, which are the history of the interaction between a task μ and an agent π over an interface ϕ. We denote by $s_\mu(t)$, $s_\pi(t)$, $a(t)$, $z(t)$, $u(t)$ the task's state, agent's state, action, observation and score, respectively, at time $t \in \mathcal{T}$. A task can be interrupted at any time (e.g., a time limit) or can stop internally.*

---

[17]Redundant Turing machines are a possibly useful way of seeing how different (Turing) machines with virtually the same capabilities can be evaluated wrongly with an inadequate interface. Redundant Turing machines are machines that work with a special coding of the input (such as 00 for 0 and 11 for 1, and –for inputs not following the code– are either null or perhaps probabilistically dependent upon those inputs which do follow the code and which are nearest in, e.g., Hamming distance) (Dowe, 2008, sec. 0.2.7)(Hernández-Orallo & Dowe, 2010, p1514, footnote 6). For example, if we redundantly code 0 and 1 as 01 and 10 respectively, if M is a machine and M_R is its redundant counter-part, and (say) $M(100) = 0110$, then $M_R(100101) = 01101001$.

[18]For instance, if a task is the addition of two natural numbers, the task internal representation of the numbers might be binary, while the interface may convert them into, e.g., a decimal representation. Note that there is a bijection between numbers in binary representation and decimal representation. Nonetheless, any transformation, even bijective, may have some potential risks. For instance, consider the task of comparing two numbers and selecting the highest one, which can be originally represented as $\langle a, b \rangle$. This can be transformed into $\langle 1, a, b \rangle$ if $a < b$ and $\langle 0, b, a \rangle$ otherwise. This transformation is bijective, but the task may become much easier for some agents, since it only implies getting the last element in the tuple. It can be argued then that there has been some information gain in the sense of (Hernández-Orallo, 2000c), which means that there might be no additional information in terms of Kolmogorov complexity, but there might still be some additional *processing* in terms of Kt complexity (Levin, 1973). Note, however, that this is not a bijection for the domain set $\langle \{0,1\}, \mathbb{N}, \mathbb{N} \rangle$. It is only a bijection if the domain set is defined as a subset of $\langle \{0,1\}, \mathbb{N}, \mathbb{N} \rangle$ where only tuples such that $\langle 1, a, b \rangle$ if $a < b$ and $\langle 0, b, a \rangle$ otherwise are accepted. This would clearly suggest that the domain set is biased for this task, as many encryptions or decryptions would also be.

This final time is denoted by t_{end}. The performance (or final score) of agent π in the episode is given by $u(t_{end})$. The functions $s_\mu(t)$, $s_\pi(t)$, $a(t)$, $z(t)$, $u(t)$ are undefined for $t > t_{end}$.

4.3. Parametric task, task class and task distribution

Examples 1 and 2 are non-deterministic. In example 1, e.g., the numbers (lights) which are randomly generated could be set as *parameters*, and hence each task *instance* would be deterministic. Let us define the notion of parametric cognitive task:

Definition 8. *A parametric cognitive task $\mu[\bar{X}]$, where $\bar{X}$ represents a tuple of parameters $x_1, x_2, ..., x_p$, is a family of tasks such that a task instance can be obtained by giving values to these parameters.*

For instance, task μ_{nc} in example 1 can be converted into a parametric cognitive task $\mu_{nc}[x_1, x_2]$ by defining two parameters $x_1, x_2 \in \mathbb{B}^*$ with initial state $s(0) = \langle \textbf{no-rewarding}, x_1, x_2, 0 \rangle$. For example, the task $\mu_{nc}[\langle 1, 0, 0, 1, 0 \rangle, \langle 1, 0, 0, 0, 0 \rangle]$ would be an instance.Now we define:

Definition 9. *A cognitive task class is a set of cognitive tasks, i.e. any subset of $\mathcal{M}$.*

For obvious reasons, task classes are more appropriate than single tasks for evaluating an ability. In fact, tests usually make a *selection* (i.e., a sample) of tasks from a task class in order to avoid, e.g., task specialisation and rote learning. A selection (or sampling) requires a distribution.

Definition 10. *A distribution over a task class M is any discrete probability distribution p such that it assigns probabilities to tasks, where for all $\mu \in M$ we have $0 \leq p(\mu) \leq 1$ and $\sum_{\mu \in M} p(\mu) = 1$.*

An example of a probability distribution for example 1 and its parametric version could be:

$$p(\mu_{nc}[x_1, x_2]) = 2^{(-2(ord(x_1) + ord(x_2)) + 1)} \tag{2}$$

where *ord* represents the order of a tuple. Assuming that $ord(x_1) \geq 1$ and $ord(x_2) \geq 1$ we have that when the total size is 2, the probability is $2^{-3} = 0.125$ (with $2^2 = 4$ possibilities, as the tuples are binary, this makes a total mass of 0.5), when the size is 3, the probability is $2^{-5} = 0.03125$ (with $2^3 = 8$ possibilities, this makes a total mass of 0.25), etc. It is then easy to see that their sum is 1.

Proposition 2. *For every task class M and distribution p there is an equivalent distribution p' for the task domain $\mathcal{M}$.*

Proof. Just define the distribution p', where $p'(\mu) = p(\mu)$ if $\mu \in M$ and 0 otherwise. $\square$

This means that a task class can be just defined as a distribution over $\mathcal{M}$.

4.4. Difficulty

Now we move to the notion of difficulty[19] applied to cognitive tasks:

Definition 11. *A difficulty (or hardness) function is any function* $\hbar : \mathcal{M} \to \mathbb{H}$*, with* $\mathbb{H} \subset \mathbb{R}^+ \cup \{0\}$*.*

In psychometrics, the difficulty of a task is commonly calculated by the results of a *population* over that task. However, while in CTT the difficulty of an item, as the proportion of examinees passing it, depends on the population in which it is calculated, in IRT we can apply what is called "parameters invariance", which means that the parameters of the model (e.g., b in eq. 1 or any other difficulty parameter) are invariant across populations, provided it is applied on the same species. As mentioned in section 3.1, there have been a few recent approaches in IRT which have tried to derive difficulty intrinsically (Sternberg, 2000, chap. 19). This can be traced back to Simon and Kotovsky, who determined item difficulty with a model for a specific kind of sequential inductive inference task (Simon & Kotovsky, 1963). The idea was extended (with more difficulties) to other tasks (Kotovsky & Simon, 1990). This significant project, however, could not be completely fulfilled without the notion of Kolmogorov complexity, compression and the invariance theorem, as done in (Hernández-Orallo & Minaya-Collado, 1998), where the difficulty of a task was determined by (a variant of) its Kolmogorov complexity.

An intrinsic, formal way of determining difficulty is much more appropriate for exploring the entire machine kingdom since we cannot practically estimate the difficulty of each task empirically as the mean score obtained by a sample of systems from this kingdom[20]. All this suggests the definition of difficulty as a function of some parameters of the cognitive class as follows:

Definition 12. *A parametric difficulty function is a difficulty function which can be defined solely with the parameters of a parametric cognitive class, i.e. it holds that there is a function* f *such that we can express the difficulty function as* $\hbar(\mu[\bar{X}]) = f(\bar{X})$*.*

An example of difficulty function for example 1 and its parametric version would be $\hbar_1(\mu_{nc}[x, y]) = ord(x) + ord(y)$, where *ord* represents the order of a tuple. For instance, given $\mu = \mu_{nc}[\langle 1, 0, 0, 1, 0\rangle, \langle 1, 0, 0, 0, 0\rangle]$ we would have $\hbar_1(\mu) = 10$. Another possible example is $\hbar_2(\mu_{nc}[x, y]) = min(\sum_{i=1}^{m} x_i, \sum_{i=1}^{n} y_i)$ where $x = \langle x_1, x_2, ..., x_m\rangle$ and $y = \langle y_1, y_2, ..., y_n\rangle$. For this difficulty function, we would have $\hbar_2(\mu) = 1$. Finally, another example would be $\hbar_3(\mu_{nc}[x, y]) = \frac{min(\sum_{i=1}^{m} x_i, \sum_{i=1}^{n} y_i)}{max(\sum_{i=1}^{m} x_i, \sum_{i=1}^{n} y_i)}$. This difficulty function, called the "ratio smaller/larger" is commonly used in tests of numerical competence (see, e.g. Cantlon & Brannon, 2006). For this difficulty function, we would have $\hbar_3(\mu) = 1/2$.

---

[19]We call it difficulty instead of complexity, especially because in the modern views of complexity derived from information theory in general, and Kolmogorov complexity in particular, complex things are not necessarily difficult. For instance, a random string is complex (in terms of its Kolmogorov complexity) but not difficult.

[20]We can, in principle, do this, by sampling all the computable agents with a distribution, and evaluate the task on a sample of them. However, this is impractical and it may boil down to assigning difficulties directly. Also, it is not clear what distribution to use, e.g. a universal distribution (Solomonoff, 1964), a Darwin-Wallace distribution (Hernández-Orallo et al., 2011) or others.

Extending the concept of average difficulty to a task class would be very useful since we usually work with parametric tasks instead of single tasks. For instance, when we refer to the task of adding two natural numbers, it is in fact a set of tasks (a parameterised task).

Definition 13. *The average difficulty of a task class M according to a difficulty function $\hbar$ and a probability distribution p is given by:*

$$H(M, p, \hbar) = \sum_{\mu \in M} p(\mu) \cdot \hbar(\mu)$$

As we see, the difficulty of the task class depends on the distribution, since it is not the same to add, e.g., one-digit numbers as ten-digit numbers.

This is a concept which is closely related to the notion of average-case computational complexity, developed by Leonid Levin in the 1980s (Levin, 1986), which is of course related to the more general notion of average-case performance of algorithms (see, e.g., Knuth, 1973). Our concept of difficulty here is not restricted to computational complexity (computational steps). Rather, the notion of 'solving a problem' is generalised to the idea of 'performing well' in a task.

It is easy to see that if the difficulty is bounded (i.e. there exists a c such that for all $\mu \in M$ we have that $\hbar(\mu) \leq c$), then the difficulty for class M is bounded. For cases where the difficulty is not bounded, we may have convergence or not, so the derived difficulty for the class may make sense or not. For instance, for the parametric cognitive task $\mu_{nc}[x_1, x_2]$ derived from example 1, and using the probability distribution in eq. 2 and the difficulty function $\hbar_1$ seen above:

$$H_1(\mu_{nc}, p, \hbar_1) = \sum_{\mu \in \mu_{nc}} p(\mu) \cdot \hbar_1(\mu) = \sum_{\mu \in \mu_{nc}} 2^{(-2(ord(x)+ord(y))+1)} \cdot (ord(x) + ord(y))$$

Let us call $n = ord(x) + ord(y)$. Since there are 2^n tasks for each n, this leads to:

$$H_1(\mu_{nc}, p, \hbar_1) = \sum_{n=1}^{\infty} 2^n \cdot 2^{(-2n+1)} \cdot n = 2 \sum_{n=1}^{\infty} 2^{(-n)} \cdot n = 2 \cdot 2 = 4$$

For $\hbar_2$ we also have convergence, since $\hbar_2(\mu) < \hbar_1(\mu)$ for every task μ. Similarly for $\hbar_3$. In other cases, however, this may not bounded. For instance, one of the most general cases (and also popular in the area of algorithmic information theory) would be to consider all the tasks using a universal distribution for probability, and Kolmogorov complexity for difficulty. The following result shows that this appealing choice does not work.

Proposition 3. *Consider the set of all cognitive tasks $\mathcal{M}$, an enumeration on them and a coding over binary strings. We define a universal distribution $p_U(\mu) = 2^{-K_U(\mu)}$, where K_U is the Kolmogorov complexity using U, a prefix UTM. And let us consider the difficulty of a task as K_U. Then, for every U we have that $H(\mathcal{M}, p_U, K_U)$ diverges.*

Proof. We can decompose the sum by the size of the tasks in the enumeration, denoted by $l(\mu)$:

$$H(\mathcal{M}, p_U, K_U) = \sum_{\mu \in \mathcal{M}} 2^{-K_U(\mu)} \cdot K_U(\mu) = \sum_{n=1}^{\infty} \sum_{\mu : l(\mu)=n} 2^{-K_U(\mu)} \cdot K_U(\mu)$$

23

Elaborating on example 4.3.4 in (Li & Vitányi, 2008, page 287), we have that there are at least $2^{(n-1)}$ tasks μ with $l(\mu) = n$ with $n - 1 \leq K_U(\mu) \leq n + 2\log n + O(1)$, so:

$$H(\mathcal{M}, p_U, K_U) \geq \sum_{n=1}^{\infty} 2^{(n-1)} \cdot (2^{-n}\frac{1}{n^2}2^{-O(1)}) \cdot (n-1) = 2^{-O(1)}\sum_{n=1}^{\infty}\frac{n-1}{2n^2}$$

By divergence of the harmonic series, the last sum diverges. □

Despite the previous negative result, we can still use Kolmogorov complexity as a measure of difficulty and still derive a proper difficulty function for task classes in general. This can be done by just defining $\hbar(\mu) = 1 - 2^{-K_U(\mu)}$ or any other bounded and monotonic transformation over K_U. Similar questions can be posed with other complexity measures, such as Levin's Kt, sophistication or logical depth (see Li & Vitányi, 2008), which may be more appropriate for measuring difficulty in general. In case of doubt of divergence, a monotonic transformation to a bounded difficulty function can be done. In fact, any of these choices would be a much more comprehensive and general way of assigning difficulty to tasks. This contrasts with ad-hoc difficulty functions for each task class, as has been generally done in AI and eventually in animal cognition and psychometrics. For instance, (Zatuchna & Bagnall, 2009) use an ad-hoc measure of difficulty for mazes. We argue that $\hbar(\mu) = 1 - 2^{-Kt(\mu)}$, where Kt is Levin's complexity (Levin, 1973) is a good option. This has been discussed and developed in (Hernández-Orallo & Minaya-Collado, 1998; Hernández-Orallo & Dowe, 2010) for inductive inference tasks, but can be extended for other kinds of tasks as well (Hernández-Orallo, 2000b,c).

5. Performance indicators: average scores and decompositions

We have not fully investigated the relation among (a) the probability of a task, (b) its difficulty, and (c) the weight we assign to each task in an overall measure. Different approaches in measurement make different choices on any of these issues, from CTT to IRT (e.g., CAT) in psychometrics to new intelligence tests in AI. In what follows, we analyse how these three issues are necessary for a proper aggregation of results and, as we will see, for an appropriate test design.

5.1. Expected average score

Given a task class M using the same interface ϕ, and a probability distribution p, we can just generate tasks, evaluate agents and calculate the average score as follows:

Definition 14. *The expected average score for a task class M, a distribution p, an interface ϕ and an agent π is:*

$$\Psi(M, p, \pi) \triangleq \lim_{\tau \to \infty} \sum_{\mu \in M} p(\mu) \cdot \mathbb{E}(Score(\pi, \mu, \phi, \tau)) \tag{3}$$

where $\mathbb{E}(Score(\pi, \mu, \phi, \tau))$ returns the expected value of the score of evaluating π with μ using interface ϕ and time limit τ.

By sampling over M using the distribution p and using finite time limits, we can approximate the above measure. The problem of the above definition is that it ignores task difficulty, and the measure (and the sampling) will be dominated by the probability distribution. This is, in principle, natural, since the performance for tasks with very low probability should be less important (for an average or aggregated score) than the performance for very frequent tasks[21]. However, when approximating the overall score, many tasks with very high probability can be very poor discriminators. This, in fact, works well in cases where the distribution is chosen in such a way that there is no small subset of tasks which dominates the distribution. But if this small subset exists, we may have agents which specialise on this subclass, or we may have that many agents perform well on this subset. This is what happens, e.g., with the intelligence measure in (Legg & Hutter, 2007), which uses a universal distribution, and a very small (and typically easy) set of tasks often dominates the distribution, as pointed out by (Hibbard, 2009) and (Hernández-Orallo & Dowe, 2010).

5.2. Decomposition

A different approach is to calculate partial results for a given difficulty h as follows:

$$U(h, M, p, \pi) \triangleq \lim_{\tau \to \infty} \sum_{\mu \in M, \hbar(\mu) = h} p(\mu|h) \cdot \mathbb{E}(Score(\pi, \mu, \phi, \tau))$$

And now, we can aggregate performance for a range of difficulties, e.g., as follows:

Proposition 4. *The expected average score* $\Psi(M, p, \pi)$, *assuming difficulty is bounded between* h_{min} *and* h_{max}, *can be rewritten as follows (if h is in a discrete domain $\mathbb{H}$, e.g., a segment in $\mathbb{Q}$ or $\mathbb{N}$):*

$$\Psi(M, p, \pi) = \sum_{h=h_{min}}^{h_{max}} U(h, M, p, \pi) p(h)$$

where $p(h)$ is the probability of a task of being of difficulty h. For continuous domains:

$$\Psi(M, p, \pi) = \int_{h_{min}}^{h_{max}} U(h, M, p, \pi) f(h) dh$$

where $f(h)$ is the probability density function of a task being of difficulty h.

---

[21]If we have a difficulty function (apart from the distribution), we may have the temptation of redefining the above eq. (3), using the difficulty as a weight, e.g., using $p'(\mu) = \frac{p(\mu) \cdot \hbar(\mu)}{\sum p(\mu) \cdot \hbar(\mu)}$, assuming the difficulty is bounded. With this new distribution, very difficult tasks would be much more frequent. In practice, however, it would fail in evaluating agents well when the sample over M, the time limits or their ability are small.
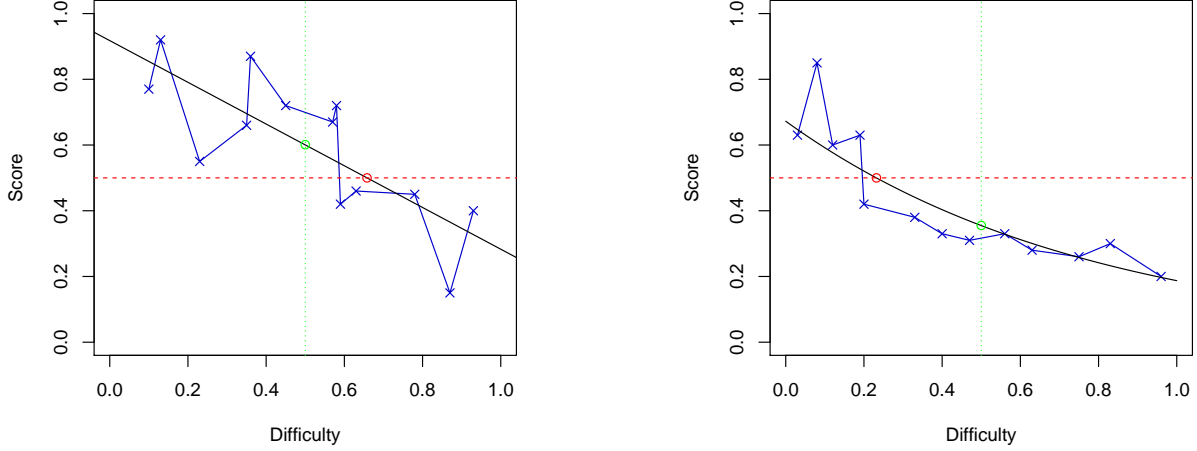
Figure 5: Two figurative Agent Response Curves for an agent and a set of of items (tasks). Results are shown in blue. Left: the results seem to fit a linear model (in black). Right: the results seem to fit a log-linear model (in black). In both cases, we show $\mathbb{E}(\hat{u}(0.5))$ (the expected score for difficulty 0.5) with a green circle and $\mathbb{E}(\hat{h}(0.5))$ (the expected difficulty for score 0.5) with a red circle.

Proof. Let us show the result for the discrete case:

$$
\begin{aligned}
\Psi(M, p, \pi) &= \lim_{\tau \to \infty} \sum_{\mu \in M} p(\mu) \cdot \mathbb{E}(Score(\pi, \mu, \phi, \tau)) \\
&= \lim_{\tau \to \infty} \sum_{h=h_{min}}^{h_{max}} \sum_{\mu \in M \, s.t. \hbar(\mu)=h} p(\mu) \cdot \mathbb{E}(Score(\pi, \mu, \phi, \tau)) \\
&= \sum_{h=h_{min}}^{h_{max}} p(h) \lim_{\tau \to \infty} \sum_{\mu \in M \, s.t. \hbar(\mu)=h} p(\mu|h) \cdot \mathbb{E}(Score(\pi, \mu, \phi, \tau)) \\
&= \sum_{h=h_{min}}^{h_{max}} U(h, M, p, \pi) p(h)
\end{aligned}
$$

A similar derivation can be done for the continuous case. $\qquad \square$

The decomposition suggests which kind of sampling can be more appropriate for estimating Ψ. Nonetheless, the most interesting application of this transformation is that we can analyse and plot the results easily, e.g., using a uniform distribution of difficulties on the x-axis.

For simplicity, we will use notation $\hat{u}(h)$ for any estimation of $U(h, M, p, \pi)$. We also define $\hat{h}(u) = \hat{u}^{-1}(u)$. As a concrete example of these concepts, Figure 5 shows the results of two different agents on the same task class. We call these curves Agent Response Curves (ARC), following similar names in the area of psychometrics, and IRT in particular (e.g., (Sternberg, 2000, chap. 19) use the term *Person Characteristic Curves*). If $p(h)$ is uniform, $\Psi(M, p, \pi)$ is just the area under the curve (i.e., the integral), either using the discrete real 'curve' (e.g., the blue points) or a continuous (parametric) curve (e.g., the black curve).
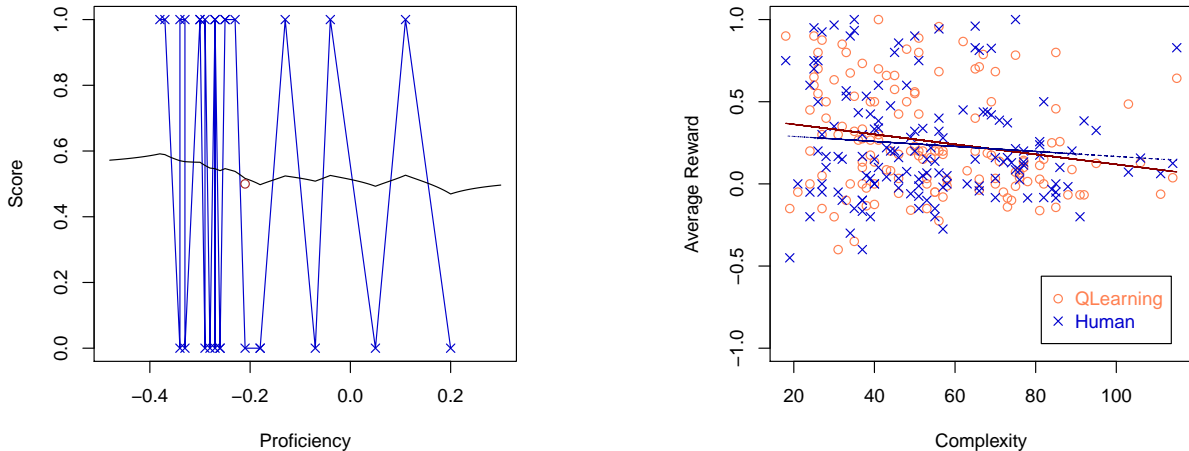
Figure 6: Left: a typical result of a CAT test (freely adapted from (Weiss, 2011, Fig. 8)). The black curve shows a Euclidean kernel smoothing with a constant of 0.1. The final proficiency calculated by the test was -0.21. Right: a comparison of performance (individual results and linear fit) of two different kinds of agents for the same task class using a difficulty function (here shown as 'complexity'), defined as an approximation of the Kolmogorov complexity of the tasks (originally published in (Insa-Cabrera et al., 2011, Fig. 4)).

Many more things can be observed on each of these figures, and some other indicators can be obtained. For instance, apart from the area (i.e., $\Psi(M, p, \pi)$), we can distinguish two interesting indicators. In Figure 5, the first point (the green circle) is where the difficulty $h = 0.5$, i.e. $\mathbb{E}(\hat{u}(0.5))$. Assuming that 0.5 is a *medium* difficulty, this point is the MDP (*medium-difficulty performance*) point (or the width at this point, in Thorndike's terminology (Thorndike, 1927)). Similarly, a second point (the red circle) is set where the score $u = 0.5$, i.e. $\mathbb{E}(\hat{h}(0.5))$. Assuming that 0.5 is a *medium* score and also a monotonically decreasing curve, this point is the maximum achieved proficiency (difficulty) with at least medium score, or simply, the MSP (*medium-score proficiency*) point. This latter point is usually known as the "point of inflection" in IRT theory (see, e.g. (Sternberg, 2000, chap. 19)), or the altitude, in Thorndike's terminology. Neither the area nor any of these points can be taken as a complete descriptor of how an agent performs. We can find examples of agents with similar area under the ARC curve, but very different indicator points.

In general, things can even be more elaborate, especially when scores are binomial or may have a high dispersion. Figure 6 (left) shows a result of the evaluation of an agent using CAT (Wainer, 2000) (the example is adapted from (Weiss, 2011, Fig. 8)). Since items can only be either correct or incorrect, the curve is a zigzag for which the calculation of an area is much more unreliable, if done as an integral, even using a smoothing (here a Euclidean kernel smoothing with a constant of 0.1 is shown). The final proficiency calculated by the CAT test in this case is -0.21 (shown as a brown circle), which may be comparable to red circles (the MSP, *medium-score proficiency* point) in figure 5. Note that, in this case, the MSP is very difficult to infer from the curve. The iterative character of CAT based on item response theory (IRT) is aimed at finding final proficiency with high reliability with the fewest possible number of items (i.e., evaluations). This proficiency represents the difficulty level such that probability of guessing the item right at this difficulty is 0.5.

Figure 6 (right) shows the results of two different agents on the same task class using a
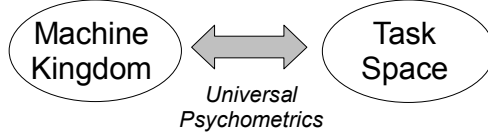
Figure 7: Universal Psychometrics can focus the exploration on any side of the duality.

non-adaptive task generator. Even though score is not binomial in this case, the plot shows how difficult it is to find the indicators (such as the area[22] or the points), since the data shows a high dispersion.

All this suggests that the way in which items are selected will depend on the kind of tasks (score domains), the range of difficulties, the expected dispersion, etc. Also, everything will be different depending on which indicators (or the whole curve) we are interested in. For instance, if scores are binomial and we want to estimate the MSP (*medium-score proficiency*) point, then the CAT approach has developed a robust theory of adaptive tests based on IRT. If scores are ordinal and we want to estimate the agent response curve, then other techniques might be useful, such as a distribution-driven Monte Carlo sampling.

In addition, for some indicators, the distribution of difficulty is important (the area), while for others it is not so (MDP and MSP). Nonetheless, the distribution for each complexity, i.e. $p(\mu|h)$, is always important for estimating the values for each h well.

6. Exploring the machine kingdom

In the previous sections we have set the fundamental pieces for exploring the machine kingdom Ω, on the one hand, and the space of abilities $\mathcal{M}$, on the other hand. This *duality* is illustrated in figure 7. The exploration of these two spaces is precisely the goal of Universal Psychometrics. This and the following sections will only highlight some basic ideas, inspired by the disciplines which have been mentioned in section 3. We will also see that while some well-established practices in these disciplines are still valid for universal psychometrics, some others are not, so alternatives need to be found.

6.1. Cognitive tests

Taking into account that we might use different indicators depending on the kind of agents, the kind of tasks and different measuring methodologies, we need to define a very general test algorithm which can fit all these situations.

Definition 15. *A Cognitive Test for a task class M and a difficulty function ℏ for that class, is a series of evaluations over tasks in M applied to an agent π (and an interface set Φ for each task in class M) as follows:*

    1. `ALGORITHM: Testing Algorithm`
    2. `INPUTS: M (a task class), Φ (interfaces), π (an agent), ℏ (a difficulty function)`
    3. `OUTPUTS: a set of indicators`
    4. `BEGIN`

---

[22]In addition, the difficulty (here shown as 'complexity') seems to be unbounded in this case, so it needs a bounded monotonic transformation on the x-axis to make the area meaningful.

5.	$p_1 \leftarrow p$	(Initial task class distribution)
6.	$\tau_1 \leftarrow \tau$	(Initial allotted time for the task)
7.	$i \leftarrow 1$	
8.	REPEAT	
9.	$\mu_i \leftarrow ChooseTask(M, p_i)$	(New task)
10.	$\phi \leftarrow ChooseInterface(\Phi, \mu_i, \pi)$	(Choose appropriate interface)
11.	$(u_i, t_i) \leftarrow Eval(\pi, \mu_i, \phi, \tau_i)$	(Returns score and actual time)
12.	$p_{i+1} \leftarrow UpdateDist(i, M, \hbar, \bar{u}, \bar{t}, \bar{\tau}, \bar{p}, \bar{\mu})$	(New distribution)
13.	$\tau_{i+1} \leftarrow UpdateTime(i, \bar{u}, \bar{t}, \bar{\tau}, \bar{\mu})$	(New allotted time)
14.	$i \leftarrow i + 1$	
15.	UNTIL $(Stoppingcriterion(i, \bar{u}, \bar{t}, \bar{\tau}, \bar{\mu}))$	(Until stopping criterion is met)
15.	RETURNS $Indicators(i, \bar{u}, \bar{t}, \bar{\tau}, \bar{\mu})$	(Returns a set of indicators)
16.	END ALGORITHM	

where $ChooseTask(M, p_i)$ chooses a task for class M using distribution p_i, $ChooseInterface(\Phi, \mu_i, \pi)$ just chooses the appropriate interface in Φ for μ_i (and π), $Eval(\pi, \mu_i, \phi, \tau_i)$ performs an episode of π for task μ_i with time limit τ_i, using the interface ϕ, $UpdateDist(i, M, h, \bar{u}, \bar{t}, \bar{\tau}, \bar{p}, \bar{\mu})$ updates the probability, $UpdateTime(i, \bar{u}, \bar{t}, \bar{\tau}, \bar{\mu})$ updates the time limits, $Stoppingcriterion(i, \bar{u}, \bar{t}, \bar{\tau}, \bar{\mu})$ decides whether to stop or go on, and, finally, $Indicators(i, \bar{u}, \bar{t}, \bar{\tau}, \bar{\mu})$ returns a set of indicators.

The generality of the algorithm is achieved by allowing all these functions to parameterise depending on each case. For instance, classical tests in psychometrics and animal cognition would just be implemented by using a distribution with all the mass on one task initially, and $UpdateDist$ would switch to another task, independently of the previous agent's result (no feedback, no adaptation). Some tests in artificial intelligence can also be in this category.

Adaptive tests in the area of CAT would also use a distribution setting where all the mass would be assigned to the task which is the most informative one, according to the statistical technique in use (maximum likelihood, other classical or Bayesian estimators — such as MML (Wallace & Boulton, 1968; Wallace & Dowe, 1999; Wallace, 2005)), in order to better (and sooner) estimate proficiency. The stopping criterion typically relies on the number of iterations or on whether this estimate is reliable. Finally, the anytime intelligence test defined in (Hernández-Orallo & Dowe, 2010) can also be considered an instance of definition 15.

6.2. Customised vs. universal tests

A general principle in measurement is that "measurement is more difficult and less efficient the less we know about the subject". If we have information about the examinee, then we can customise the test in order to have a more effective and efficient measurement. That is the reason why several disciplines have customised tests for adult humans, for children, for people with disabilities, for different species and for different kinds of AI algorithms.

However, there are two reasons why customised tests are inappropriate for exploring the machine kingdom. On the one hand, customised tests do not allow the straight comparison between different kinds of agents in an even way, since only with the same task selection procedure, the same indicators and the same stopping criteria we can reliably compare results. On the other hand, we may have no (or very little) information about the subject

we want to evaluate —and this may be the case when evaluating an unknown machine, and will be much more frequent in the future. In any of these situations we need a test that is valid for any kind of individual.

Definition 16. *A universal test for cognitive task class M is a test which is able to measure the ability on M for any element in the machine kingdom Ω.*

The emphasis is set on "universal". If a cognitive test is useful for measuring one cognitive ability for a group (kind, series or species) of systems but useless —or inaccurate— for other groups, then the test is not universal. Conversely, and following the (physical) Church-Turing thesis (Copeland, 2008), if a test is able to measure a cognitive ability for any conceivable machine, then, by the Church-Turing thesis, it will be able to measure this cognitive ability for animals (and as a consequence, for humans).

The design of universal tests is much more difficult than that of customised tests. One requirement for a universal test is that it has to be necessarily *adaptive*, since we do not know in advance the range of difficulties and the time resolution the agents can take. This is, for instance, the approach taken in (Hernández-Orallo & Dowe, 2010), which has been further generalised here in definition 15.

A universal test requires a choice of interfaces beforehand. The important thing is that the task definitions and the testing procedure are the same. For instance, (Herrmann et al., 2007) uses the same tasks for human children and adult chimpanzees by carefully changing the interfaces. Similarly, in (Insa-Cabrera et al., 2011), humans and reinforcement learning (RL) algorithms are compared on the same task generator but a different interface.

Clearly, universal tests are much more difficult to construct than customised tests (see Table 1), but they are necessary for exploring the machine kingdom. However, customised tests are still useful for practical and theoretical purposes, provided there is some degree of overlapping in the population, in order to make comparisons and alignments of results. In the end, universal psychometrics has to work with both types of tests: (population-)customised and universal.

6.3. Populations, normalisation and scaling

The use of a population of agents is the traditional approach in psychometrics. In the machine kingdom, instead of using an empirical population, we can use a theoretical population. In order to do that, we need, as well, a distribution over agents.

Definition 17. *A population of agents Π is any subset of Ω with an associated probability distribution ρ from Ω.*

Expected average score (definition 14) was presented for a single individual, agent or subject, but we can extend it for populations.

$$\Psi(M, p, \Pi, \rho) \triangleq \lim_{\tau \to \infty} \sum_{\mu \in M} p(\mu) \sum_{\pi \in \Pi} \rho(\pi) \cdot \mathbb{E}(Score(\pi, \mu, \phi, \tau)) \qquad (4)$$

The decomposition of $\Psi(M, p, \pi)$ (definition 14) according to proposition 4 can also be applied to the above formula, but other decompositions can be figured out, e.g., using the capabilities of the agents. In fact, in general, the use of agent response curves, such as those

Feature	Population-customised test	Universal test
Difficulty	Empirical estimation (e.g., IRT)	Theoretically-derived, based on task class
Administration	Sequential or adaptive	Necessarily adaptive
Discrimination	Empirical estimation (e.g., IRT)	Theoretically-derived, based on item diversity
Validity	Models of cognitive abilities, factor analysis, content analysis, significance tests	Formal (computational) models of cognitive abilities
Scaling	Empirical normalisation	Theoretically-derived, based on task class characteristics
Analysis of results	(Latent) factor analysis, significance tests	Significance tests
Conditioning	Instruction, final reward	Rewards

Table 1: Comparison of measurement features. The table shows how measurement features can be typically addressed in customised tests for a well-known population (e.g. humans) vs. universal tests.

depicted in figures 5 and 6, is more frequently applied to compare populations (or species) rather than comparing specific individuals (see, e.g., page 431 in Sternberg, 2000). Figure 6 (right) is an example of this.

The use of populations (either empirical or theoretical) may suggest their use for normalising or scaling scores and difficulties. For instance, one option for an 'absolute' scaling would be to select all the agents in Ω and set this average as a zero point. This would be a generalisation of what is done for IQ tests, which use a Gaussian distribution with mean at 100 and standard deviation 15 (or 16, in some tests), over a population. However, how can we sample from Ω, the set of all agents? We may suggest using a universal distribution over Ω. However, this would be very sensitive to the reference machine used for the universal distribution. Also, for most reference machines, the agents with shortest descriptions would probably have very anomalous behaviours (such as not reacting at all, as a 'dead' agent). A different approach would be to *scale* the scores by using a reference or baseline machine, chosen to score, e.g., 0. This can be done as follows.

When time between actions is not considered in the measurement and the set of actions $\mathcal{A}$ is finite, then a proper reference agent is a random agent π_{rand}, as an agent which chooses uniformly in $\mathcal{A}$. This is the approach taken in (Hernández-Orallo & Dowe, 2010), where tasks (and scores) are chosen in such a way that:

$$\forall \mu, \tau \quad \mathbb{E}(Score(\pi_{rand}, \mu, \phi, \tau)) = 0 \tag{5}$$

where scores range between -1 and 1. Eq. 5 implies that (but it is not a necessary condition for) $\Psi(M, p, \pi_{rand}) = 0$. The notions of performing worse than random or better than random are meaningful and appropriate. This is similar to the statistical correction in multiple choice questions, where the result is corrected by the probability of guessing it right by chance. When $\mathcal{A}$ is infinite, the notion of random agent becomes more elusive, since we cannot use a uniform distribution.

When time is considered, the choice of a reference agent is much more difficult, since there are many possibilities of how time can be considered (and its relation to performance, see, e.g., Hernández-Orallo, 2010). The use of a constant-pace random-action agent is not a good reference either, since we need to arbitrarily determine an action rate. In general, there is no reference agent that can be used for all tasks. However, for any class task, we may be able to find a reasonable reference agent and scale the bottom of the score accordingly.

6.4. Evaluating development: potential vs. actual intelligence

The notions of potential versus actual abilities are usually linked to the 'nature' vs. 'nurture' debate (Eysenck, 2007). What can we say in general for non-human animals and machines? Several efforts have been made to bring up apes in a human context. While the results show that some improvement can be shown because of the enriched human context, animal abilities are limited by their genes. What about machines? Things are more complex here. Theoretically, it is possible to construct a machine such that it is unintelligent until an appropriate "training" signal is received, when it changes into another state where it becomes intelligent (Solomonoff, 1962). In fact, for any universal Turing machine (UTM) there is an input —a program— such that the machine becomes any particular other machine, e.g., a machine with any degree of actual intelligence. So we could loosely say that any UTM has maximal potential intelligence (and the same applies for any cognitive ability). However, if we construct a second machine such that this second state is accessed much more easily (without the need of a very specific input), we can intuitively say that the second machine has more potential intelligence.

This suggests an indirect way of measuring "potential" intelligence (and other abilities). This can be done by using distributions of "lives" (or training environments). The distribution and life span of the environment may be used to parameterise the definition.

Naturally, potential abilities (e.g., potential intelligence) are very difficult to measure but this issue cannot be ignored[23], since it is quite unlikely that we can construct an algorithm such that it makes a machine intelligent the first day. Surely, the machine will require some training. How difficult the training is will of course be related to potential abilities. This raises of course the idea that the goal of artificial intelligence is perhaps to construct potential intelligence rather than actual intelligence, which is closely related to the construction of systems that learn. Creating an "AI baby" has been an idea that is not new in the discipline (Turing, 1950; Solomonoff, 1962; Goertzel & Bugaj, 2009). If this is finally the right track in AI, then universal psychometrics will be absolutely necessary, and inductive inference abilities will play a central role in the analysis of potential versus actual intelligence. For a more extensive account of the notion of potential intelligence for machines, we refer to (Hernández-Orallo & Dowe, 2013).

---

[23]This is also important ethically, since our notion of "person" is related to "potential" intelligence rather than "actual" intelligence. Babies have human rights, some years before they can show much "actual" intelligence.

7. Exploring the space of abilities

The relation between abilities is one of the most debated issues in psychometrics, where many different models have been proposed, as we saw in section 3. Exploring the space of abilities (or, more precisely, the space of tasks, see figure 7) in general (i.e., in the machine kingdom) seems even more challenging and controversial.

7.1. Validity, reliability and relations among abilities

Validity, in the most classical meaning, can be expressed as the correspondence between a set of tasks and the ability it is supposed to measure. How can we choose a representative set of tasks for an ability? We can find the correspondence between abilities (or other constructs, such as factors), and the instruments (the task classes and the tests) using empirical techniques (such as factor analysis) over a population. However, any conclusion we may reach will only be applicable to that population. For instance, if we find that numerical abilities and spatial abilities correlate in humans, that does not mean that this should also be the case for, e.g., chimpanzees, or other machine subpopulations. On the other hand, an empirical approach using the whole set of agents $\mathcal{M}$ seems clearly unrealistic. A theoretical approach seems, then, the alternative.

Theoretical approaches in psychometrics exist and validation is seen as the result of accumulating evidence supporting the theoretical model and "to provide a sound scientific basis for the proposed score interpretations" (AERA, APA & NCME, 1999). However, tasks and abilities are not mathematically formalised. Nonetheless, a mere formalisation of a task does not yield its full potential *unless* it is derived from a meaningful definition of a cognitive ability. For instance, inductive inference has been rigorously formalised while still corresponding to its intuitive notion. This can be found in Solomonoff's theory of prediction (Solomonoff, 1964) (using a mixture of models) and the theory of inductive inference based on the Minimum Message Length (MML) principle (Wallace & Boulton, 1968) (using just one model, obtained by two-part compression), both related to Kolmogorov complexity (Wallace & Dowe, 1999; Li & Vitányi, 2008). In (Hernández-Orallo & Minaya-Collado, 1998; Hernández-Orallo, 2000a) these formal theories of inductive inference are adapted to generate exercises of inductive inference ability. With this, the validity of the tests has a better foundation from the start, and we have more tools to assess difficulty, to derive measures or conditions for discrimination, and, in the end, to derive tests which can get higher reliability.

The same can be done for other general or specific abilities, such as deductive abilities or rote memory, which seem easier, while spatial abilities and verbal abilities seem more difficult (but not impossible) to define in a formal way. In (Hernández-Orallo, 2000b) some of these abilities are defined based on the notion of information gain (Hernández-Orallo, 2000c), which allows for the derivation of "difficulty" for deductive problems as well, also using Kolmogorov complexity. Using this theory, there are some proposals for more challenging abilities, such as creativity (Dowe, 2008; Schmidhuber, 2010, sec. 0.2.7, p545, col. 1)(Dowe, 2011, sec. 7.7). Extending the set of abilities is necessary if we do not want to limit universal psychometrics to intelligence —or a very small facet of it. Some experiments performed with the first computational definitions and tests of intelligence have shown that some non-intelligent systems scored competitively when compared to humans in some of these tests (Insa-Cabrera et al., 2011). Although we can look for the reason in how the tests were

derived or administered, we have to acknowledge that we have also seen this phenomenon for custom IQ tests (Sanghi & Dowe, 2003). In any case, this still means that intelligence testing (and, needless to say, other abilities) must consider broader definitions and sets of tasks than those considered to date with the computational approach.

A mathematical and precise definition of each ability is not only important for validity and reliability but also for finding relations between abilities. This approach configures correlation (or, more properly, relation) as an absolute and not a relative property. For example, without a proper formalisation of abilities, it is very difficult to know whether some abilities correlate because they go together with the species (or even with all mammals) or rather because they must go together. For instance, it is well known that deductive abilities are required for inductive abilities. Consistently, they correlate in humans and perhaps this holds in general.

This is indeed related to the idea of falsifiability. Consistently finding that two cognitive abilities always correlate for a certain species (or even for many species) does not prove that the two cognitive abilities are similar. There is always the possibility that a new species (or individual), or a machine, can show a behaviour where these two abilities do not go together.

7.2. General abilities, intelligence and no free lunch

Another question is whether some complex cognitive abilities can be defined from simpler ones, as done typically in a hierarchical way in psychometrics. Again, the hierarchy should be derived theoretically rather than empirically. If we look at AI, though, it is difficult to find a similar thing. The closest notion where we can see several distinguished abilities is when we take a look at the table of contents of any AI handbook. There we usually find a "taxonomy", where each chapter corresponds to subdisciplines in AI such as machine learning, automated deduction, planning, natural language processing, pattern recognition, perception, multi-agent systems, social intelligence, creativity, etc. It is easy to find a correspondence between these disciplines and the abilities usually recognised in psychometrics. Nonetheless, in AI there is no such a thing as a uniform measuring methodology for all of them. In fact, there is no consensual methodology in any of them separately.

Some models in psychometrics have suggested a hierarchical view of cognitive abilities and/or factors (Jensen, 1998)(Sternberg, 2000, chap. 3), where general intelligence is typically at the top, with more specific abilities at a second (or third) level. From this hierarchical view and the definition of cognitive task classes as given here, we can suggest the notion of the most general ability as *performance in all tasks*. In our context, this boils down to performance in $\mathcal{M}$, the set of all possible tasks. The *question* is how the distribution over this set is chosen. As we know, this cannot be chosen uniformly, because $\mathcal{M}$ is a recursively enumerable infinite set of tasks. In addition, some other 'regular' choices give problems. For instance, distributions cannot be symmetrical in terms of scores:

Proposition 5. *Assume that we have a probability distribution p over $\mathcal{M}$ giving the same probability to any task $\mu^{\oplus}$ in $\mathcal{M}$ and its 'complementary' task $\mu^{\ominus}$ in $\mathcal{M}$, which is exactly equal to $\mu^{\oplus}$ except that $\dot{u}_1 = -\dot{u}_2$. Under this assumption, any agent performs equally with expected average score 0.*

Proof. Consider that agent π scores σ on task μ. Since $\mathcal{M}$ contains all tasks, there is a task μ' such that μ and μ' are complementary. Since scores are complementary, it must score $-\sigma$.

Hence:

$$\Psi(\mathcal{M}, p, \pi) = \lim_{\tau \to \infty} \sum_{\mu \in \mathcal{M}} p(\mu) \cdot \mathbb{E}(Score(\pi, \mu, \phi, \tau))$$

can be decomposed as $\mathcal{M} = \mathcal{M}^{\oplus} \cup \mathcal{M}^{\ominus} \cup \mathcal{M}^0$ (with $\mathcal{M}^{\oplus} \cap \mathcal{M}^{\ominus} = 0$ and $\mathcal{M}^0$ being the set of tasks where $\dot{u}$ is everywhere 0). Then,

$$
\begin{aligned}
\Psi(\mathcal{M}, p, \pi) &= \Psi(\mathcal{M}^{\oplus}, p^{\oplus}, \pi) + \Psi(\mathcal{M}^{\ominus}, p^{\ominus}, \pi) + 0 \\
&= \lim_{\tau \to \infty} \sum_{\mu \in \mathcal{M}^{\oplus}} p^{\oplus}(\mu) \cdot \mathbb{E}(Score(\pi, \mu, \phi, \tau)) + \lim_{\tau \to \infty} \sum_{\mu \in \mathcal{M}^{\ominus}} p^{\ominus}(\mu) \cdot \mathbb{E}(Score(\pi, \mu, \phi, \tau))
\end{aligned}
$$

Since the probabilities are the same, $p^{\oplus} = p^{\ominus}$ and scores symmetrical and bounded (see definition 5), this leads to:

$$\Psi(\mathcal{M}, p, \pi) = \lim_{\tau \to \infty} \sum_{\mu \in \mathcal{M}^{\oplus}} p^{\oplus}(\mu) \cdot \mathbb{E}(Score(\pi, \mu, \phi, \tau)) - \lim_{\tau \to \infty} \sum_{\mu \in \mathcal{M}^{\oplus}} p^{\oplus}(\mu) \cdot \mathbb{E}(Score(\pi, \mu, \phi, \tau)) = 0$$

Then for every π we have that $\Psi(\mathcal{M}, p, \pi) = 0$. $\qquad\square$

The previous case is just a possible situation where discriminative *power* is zero. For example, there are other distributions where all agents always perform equally

Proposition 6. *If a distribution over $\mathcal{M}$ gives probabilities to the score functions independently from the rest of the task, i.e., there are two probability distributions p_u and p_t such that for all μ (defined as the tuple $\langle \mathcal{T}, \mathcal{S}, \mathcal{O}, \mathcal{I}, \dot{s}, \dot{o} \rangle$ and a score function $\dot{u}$), we have that:*

$$p(\mu) = p_u(\dot{u}) \times p_t(\langle \mathcal{T}, \mathcal{S}, \mathcal{O}, \mathcal{I}, \dot{s}, \dot{o} \rangle)$$

then any agent performs equally.

Proof. Here we just need to see that scores are independent of what the agent does (on average). $\qquad\square$

Fortunately, the previous proposition does not hold for any subset of $\mathcal{M}$ such that rewards are used and they are *positively correlated* with final score. All this implies that task classes must be chosen in such a way that scores must be meaningful and related to the task.

The previous results are closely related to the "no free lunch" theorems, whose more general (and less technical) version state that "any two algorithms are equivalent when their performance is averaged across all possible problems" (Wolpert & Macready, 2005), while the average uses "any distribution over objective functions which is invariant under permutation of the space of candidate solutions". In proposition 6 we have a related version of invariance for score functions and in proposition 5 we can see a restricted view of permutation, with similar consequences.

The no free lunch theorems can also be applied to specific abilities and not to the overall mixture of tasks. For instance, if we take the task class of predicting binary sequences, if all sequences are equally probable (or, more properly, if probabilities are preserved after partial permutation) then any evaluable system will get the same result. In fact, this is

interesting, because if we do that, i.e., if the no-free-lunch theorems were of applicability, then no IQ test would be valid, since any answer in a test would be equally reasonable. This is the "subjectivity objection" again (see section 3.1), i.e., the plausibility of 9 following 1,3,5,7 would be the same as 1 following it. And, in the end, measurement of many abilities covering an infinite number of possibilities would be simply impossible.

Fortunately, the assumption that given a set of tasks representing an ability all should be equally probable is not sustainable. On the one hand, when these sets are infinite the probability is zero. On the other hand, it is not true that all problems are equally probable.

This leads to the notion of universal distributions derived from Kolmogorov Complexity, known as the Solomonoff prior. Although this option is not the only possible way of assigning non-zero probabilities to an infinite set, the use of this kind of prior has several reasonable properties (Li & Vitányi, 2008). In some way or another, most of the approaches based on information theory in Section 3.3 have used this distribution. In fact, some recent results by Hibbard (2009) show "rigorous support that an intelligence measure must be based on unequal weighting of environments, such as the weighting based on Kolmogorov complexity".

However, universal distributions are not immune to low or zero discriminative power. For tasks, for example, there might be universal distributions (depending on the machine) for which proposition 5 can be applied. For instance, consider a machine such that the first bit of the program is always used to say whether the scores are negated or not. For each task we would have an equivalent task with the symmetrical score result and equal description size (and equal algorithmic probability).

This does not happen if we use meaningful rewards and performance is defined as an average of rewards. So, it is the combination of non-uniform distributions with a proper assignment of scores which makes the use of the whole (or a big family in) $\mathcal{M}$ possible (and meaningful) as formalisations of very general abilities (e.g., intelligence), such as the C-test (Hernández-Orallo & Minaya-Collado, 1998), the so-called universal intelligence measure (Legg & Hutter, 2007) and the anytime intelligence test (Hernández-Orallo & Dowe, 2010).

Finally, the definition of intelligence as performance on all tasks (with a proper distribution) connects with the role of the g factor. If any cognitive ability is seen as corresponding to a subset over $\mathcal{M}$, then good performance on $\mathcal{M}$ is expected to imply good performance in significant big subsets of $\mathcal{M}$, which may represent other abilities, as happens with g[24]. We have to be cautious about this, as some early tests using algorithmic information theory have a high g-factor load but have been shown to be insufficient to evaluate the multi-faceted concept of intelligence, especially if these tests are applied to non-human animals and machines. Similarly, some (new) emergent properties can appear as the result of a combination of several subsystems lacking these properties, such as neural computing systems,

---

[24]A possible understanding of the g factor in this context of distributions is to link it with the invariance theorem (a universal Turing machine provides an optimal means of description, up to an additive constant), first proven by Ray Solomonoff (Solomonoff, 1964) and also with his other very important theorem on the bounds on prediction error (Solomonoff, 1978), which can be extrapolated (with somewhat less tight bounds) to the related –but different– problem of inductive inference (Dowe, 2011, sec. 4). It is not a surprise then, as it has been found in humans, that "inductive reasoning is the primary ability most commonly associated with g" (Sternberg, 2000, chap. 14), since this would be the ability of learning to perform well in any other ability.

multi-agent systems, swarms, multi-player games, robotic teams, social networks, etc., with varying degrees and types of connection topologies, communication protocols or languages. This is not only important to the present and future of cognitive systems and (artificial) intelligence but it is our contention that it may provide greater impetus for the unification in universal psychometrics.

8. Discussion

'Universal psychometrics', as presented so far, opens many new questions and may be subject to many objections. The most important objection is about its feasibility. We do not expect a conclusive answer to this objection, as this will be, in any case, a progressive endeavour, where better and more general theories and tools will be eventually introduced. In this paper, we have just outlined some fundamental elements and properties, which may of course be questioned and revised in the future. Nonetheless, the construction of tests for the machine kingdom cannot be eluded. The following decades will increasingly demand cognitive tests for new cognitive systems, including machines and hybrids. Once (and if) the *technological singularity* (Solomonoff, 1962; Good, 1965; Solomonoff, 1985) is achieved, this will be a necessity (Hernández-Orallo & Dowe, 2010, p1509 and p1536).

Of course we can have ethical objections. Some could say that it is too pretentious to put humans, non-human animals, machines and hybrids (and communities) thereof on the same scale, and measuring them with the same ruler. We cannot disagree more. Measuring things with the same rulers is one of the principles in science. That does not mean that it has to be one ruler but rather a set of (possibly related) rulers for many abilities. The fact that the 'scala naturae' scheme (see section 3.2) has been abandoned in evolution and in the study of cognition does not mean that we cannot have different scales for many abilities. What we argue here is that the ruler for each ability must be the same for all. In fact, focussing on intelligence alone may give a poor characterisation of the agents we want to evaluate, since it is the range of abilities and the results an agent obtains in each of them which characterises the agent.

We have clearly defined the machine kingdom as the set of all resource-bounded computable machines. There is an exceptional opportunity whenever a realm or field of study is broadened. The good thing about universal psychometrics —and the ruling principle for the discipline— is that its tests are falsifiable (in Popper's sense) by experimentation. If a test is assumed to measure an ability and we find (or construct) a subject for which the test repeatedly evaluates the ability in a wrong way, then we have falsified the test. This principle has been used regularly in psychometrics, but the scope of the machine kingdom (including virtually any possible system) makes falsifiability more powerful. For instance, for humans, it has been shown that several abilities correlate, but we do not know whether these abilities correlate theoretically or whether it is just because they happen frequently together. Evaluating these abilities on machines (which is far easier than on other animals) may help us to clarify some of these questions.

Nonetheless, it is important to state that we are not proposing here that classical and functional IQ tests should be replaced by more general "universal" tests when the goal is evaluating humans or better understanding human intelligence. Most probably, universal tests will be less effective, since they assume fewer things about the examinee. Evaluating

humans and animals has many specific issues which are difficult to address with such a general view originating from universal psychometrics, such as the notion of "species", the evolutionary background, the genetic load, the appraisal of intelligence development, etc. The view of "universal psychometrics" as a superdiscipline should be better substituted by a view of "universal psychometrics" as an overlapping or transversal discipline which would allow the exchange of theories, techniques and methodologies among some of the previously isolated disciplines. However, universal psychometrics should be mainly constructed from formal, computational, principles. In this sense, this work has presented the foundations of a general approach for universal psychometrics, with the space of evaluable systems (the machine kingdom) and the space of tasks as the field of study. We have revisited and formally redefined the notions of (expected) performance, difficulty, interface, test, etc., and we have discussed some key issues around them, as well as some results showing possible directions —and some dead-ends, too.

The general perspective of universal psychometrics enlarges both the theoretical and the experimental possibilities. In one direction, we can bring the experience, terminology, experimentation methodology, findings and expertise from psychometrics to the realm of machine intelligence measurement and to the field of AI in general. In particular, AI evaluation may receive an important impact from the view of a cognitive system being characterised by the diversity and degree of its cognitive abilities, the analysis of the relation between the spectrum of abilities, the distinction between traits and measuring instruments, or the borrowing of item response theory, as well as other theories and concepts developed in psychometrics. In the other direction, as already mentioned, we can derive tasks and problems from non-anthropocentric mathematical principles, and determine their difficulty theoretically. Similarly, as the areas of animal behaviour and cognition have become more and more related to computational disciplines such as artificial life, artificial evolution, swarm computation and biorobotics, there are many possibilities to be explored when we focus on cognitive system evaluation and ability measurement. The notion of *animat* can be crucial in this connection, as different animats need to be compared to real animals (and other animats) and the evaluation tools can benefit from comparative cognition and artificial intelligence alike. Finally, universal psychometrics can strengthen the already regular connections between psychometrics and comparative cognition, as there is a continuum in the machine kingdom between animats and androids; avatars, assistants, robots, chatterbots and robotic pets are designed and chosen because of their range of abilities and personality.

It is very difficult to tell how much time and effort it will take to get a universal intelligence test which could be applied to any machine, and, as a result of the physical Church-Turing thesis, as already discussed, be applicable to any animal, including humans, or any hybrid or collective. It is perhaps better to ask the question for some other simpler cognitive abilities, and see whether an aggregated cognitive ability called intelligence makes sense in the context of universal psychometrics. Also, in this paper we have mostly talked about cognitive abilities, but there are some other features and traits which would be useful to measure on machines, in the same way psychology and psychiatry do: empathy, aggressiveness, perseverance, etc. But this is still further away.

We can conclude referring to the main motivation already expressed by some earlier works in the 1990s about measuring cognitive systems: artificial intelligence requires an accurate, non-anthropocentric, meaningful and formal way of evaluating its progress, by evaluating its

artefacts.

Acknowledgments

References

AERA, APA & NCME (1999). *Standards for educational and psychological testing*. American Educational Research Association.

von Ahn, L., Blum, M., & Langford, J. (2004). Telling humans and computers apart automatically. *Communications of the ACM*, *47*, 56–60.

Barrera, A., Cáceres, A., Weitzenfeld, A., & Ramirez-Amaya, V. (2011). Comparative experimental studies on spatial memory and learning in rats and robots. *Journal of Intelligent and Robotic Systems*, *63*, 361–397.

Beck, B. B. (1982). Chimpocentrism: Bias in cognitive ethology. *Journal of Human Evolution*, *11*, 3–17.

Bitterman, M. E. (1960). Toward a comparative psychology of learning. *American Psychologist*, *15*, 704–712.

Boring, E. G. (1923). Intelligence as the tests test it. *New Republic*, (pp. 35–37).

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Borsboom, D., Mellenbergh, G., & Van Heerden, J. (2004). The concept of validity. *Psychological review*, *111*, 1061.

Brent, R. P. (2008). Some comments on CS Wallace's random number generators. *The Computer Journal, Christopher Stewart WALLACE (1933-2004) memorial special issue*, *51*, 579–584.

Bringsjord, S. (2011). Psychometric artificial intelligence. *Journal of Experimental & Theoretical Artificial Intelligence*, *23*, 271–277.

Bringsjord, S., & Schimanski, B. (2003). What is artificial intelligence? Psychometric AI as an answer. In *International Joint Conference on Artificial Intelligence* (pp. 887–893).

Bullock, S., & Cliff, D. (2004). Complexity and emergent behaviour in ICT systems. *Digital Media Systems Laboratory, HP Laboratories Bristol*, .

Burt, C. (1949). The structure of the mind: A review of the results of factor analysis. *British Journal of Educational Psychology*, *19*, 100–111, 176–199.

Cantlon, J. F., & Brannon, E. M. (2006). Shared system for ordering small and large numbers in monkeys and humans. *Psychological Science*, *17*, 401.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Chaitin, G. J. (1982). Godel's theorem and information. *International Journal of Theoretical Physics*, *21*, 941–954.

Copeland, B. J. (2008). The Church-Turing Thesis. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University. (Fall 2008 ed.).

Detterman, D. K. (2011). A challenge to Watson. *Intelligence*, *39*, 77 – 78.

Deutsch, D. (1985). Quantum theory, the Church-Turing principle and the universal quantum computer. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, *400*, 97–117.

Di Paolo, E. (August 2009). Editorial. *Adaptive Behavior*, *17*, 267.

Dobrev, D. (2000). AI - What is this? A definition of artificial intelligence. *PC Magazine Bulgaria (in Bulgarian, English version at http://www.dobrev.com/AI)*, .

Dobrev, D. (2005). Formal definition of artificial intelligence. *International Journal of Information Theories and Applications*, *12*, 277–285.

Dowe, D. L. (2008). Foreword re C. S. Wallace. *Computer Journal*, *51*, 523 – 560. Christopher Stewart WALLACE (1933-2004) memorial special issue.

Dowe, D. L. (2011). MML, hybrid Bayesian network graphical models, statistical consistency, invariance and uniqueness. In P. S. Bandyopadhyay and M. R. Forster (Ed.), *Handbook of the Philosophy of Science - Volume 7: Philosophy of Statistics* (pp. 901 – 982). Elsevier.

Dowe, D. L., & Hajek, A. R. (1997). A computational extension to the Turing Test. *in Proceedings of the 4th Conference of the Australasian Cognitive Science Society, University of Newcastle, NSW, Australia*, .

Dowe, D. L., & Hajek, A. R. (1998). A non-behavioural, computational extension to the Turing Test. In *Intl. Conf. on Computational Intelligence & multimedia applications (ICCIMA'98), Gippsland, Australia* (pp. 101–106).

Dowe, D. L., & Hernández-Orallo, J. (2012). IQ tests are not for machines, yet. *Intelligence*, *40*, 77–81.

Dowe, D. L., Hernández-Orallo, J., & Das, P. K. (2011). Compression and intelligence: social environments and communication. In J. Schmidhuber, K. Thórisson, & M. Looks (Eds.), *Artificial General Intelligence* (pp. 204–211). LNAI series, Springer volume 6830.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. L. Erlbaum.

Evans, T. G. (1964). *A program for the solution of a class of geometric-analogy intelligence-test questions*. Technical Report DTIC Document, also appeared later in Minsky M. (ed.) Semantic Information Processing, pp. 271-353, Cambridge, Massachussets, 1968.

Eysenck, H. J. (2007). *The Structure and Measurement of Intelligence*. Transaction Publishers.

Genesereth, M., Love, N., & Pell, B. (2005). General game playing: Overview of the AAAI competition. *AI Magazine*, *26*, 62.

Goertzel, B., & Bugaj, S. V. (2009). AGI preschool: a framework for evaluating early-stage human-like AGIs. In *Proc. of the 2nd Intl. Conf. on Artificial General Intelligence (AGI-09)*.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. *Advances in Computers*, *6*, 31–88.

Gosling, S. D. (2001). From mice to men: what can we learn about personality from animal research? *Psychological bulletin*, *127*, 45.

Griffin, D. R. (1978). Prospects for a cognitive ethology. *Behavioral & Brain Sciences*, *1*, 527–538.

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Harlow, H. F., Harlow, M. K., & Meyer, D. R. (1950). Learning motivated by a manipulation drive. *Journal of Experimental Psychology*, *40*, 228–234.

Hernández-Orallo, J. (2000a). Beyond the Turing Test. *J. Logic, Language & Information*, *9*, 447–466.

Hernández-Orallo, J. (2000b). On the computational measurement of intelligence factors. In A. Meystel (Ed.), *Performance metrics for intelligent systems workshop* (pp. 1–8). National Institute of Standards and Technology, Gaithersburg, MD, U.S.A.

Hernández-Orallo, J. (2000c). Thesis: Computational measures of information gain and reinforcement in inference processes. *AI Communications*, *13*, 49–50.

Hernández-Orallo, J. (2010). On evaluating agent performance in a fixed period of time. In M. H. et al. (Ed.), *Artificial General Intelligence, 3rd Intl Conf* (pp. 25–30). Atlantis Press.

Hernández-Orallo, J., & Dowe, D. L. (2010). Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, *174*, 1508 – 1539.

Hernández-Orallo, J., & Dowe, D. L. (2013). On potential cognitive abilities in the machine kingdom. *Minds and Machines*, *23*, 179–210.

Hernández-Orallo, J., Dowe, D. L., España-Cubillo, S., Hernández-Lloreda, M. V., & Insa-Cabrera, J. (2011). On more realistic environment distributions for defining, evaluating and developing intelligence. In J. Schmidhuber, K. Thórisson, & M. Looks (Eds.), *Artificial General Intelligence* (pp. 82–91). LNAI, Springer volume 6830.

Hernández-Orallo, J., & Minaya-Collado, N. (1998). A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. In *Proc. Intl Symposium of Engineering of Intelligent Systems (EIS'98)* (pp. 146–163). ICSC Press.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, *Vol 317*, 1360–1366.

Herrmann, E., Hernández-Lloreda, M. V., Call, J., Hare, B., & Tomasello, M. (2010). The structure of individual differences in the cognitive abilities of children and chimpanzees. *Psychological Science*, *21*, 102.

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

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica*, *26*, 107–129.

Insa-Cabrera, J., Dowe, D. L., España-Cubillo, S., Hernández-Lloreda, M. V., & Hernández-Orallo, J. (2011). Comparing humans and AI agents. In J. Schmidhuber, K. Thórisson, & M. Looks (Eds.), *Artificial General Intelligence* (pp. 122–132). LNAI, Springer volume 6830.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, Praeger.

Kelley, T. L. (1927). *Interpretation of educational measurements.*. World Book Co.

Kelley, T. L. (1928). *Crossroads in the mind of man: A study of differentiable mental abilities.*. Stanford university press.

Knuth, D. E. (1973). *Sorting and searching, volume 3 of The Art of Computer Programming*. Addison-Wesley.

Kotovsky, K., & Simon, H. A. (1990). What makes some problems really hard: Explorations in the problem space of difficulty. *Cognitive Psychology*, *22*, 143–183.

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, *17*, 391–444.

Legg, S., & Veness, J. (2011). An Approximation of the Universal Intelligence Measure. In *Proceedings of Solomonoff 85th memorial conference, Lecture Notes in Computer Science (LNCS)* (pp. 217–231). volume 7070.

Levin, L. A. (1973). Universal sequential search problems. *Problems of Information Transmission*, *9*, 265–266.

Levin, L. A. (1986). Average case complete problems. *SIAM J. on Computing*, *15*, 285–286.

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

Long, D., & Fox, M. (2003). The 3rd international planning competition: Results and analysis. *J. Artif. Intell. Res. (JAIR)*, *20*, 1–59.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores.*. Addison-Wesley.

MacLean, E. L., Matthews, L. J., Hare, B. A., Nunn, C. L. et al. (2011). How does cognition evolve? phylogenetic comparative psychology. *Animal Cognition*, (pp. 1–16).

Mahoney, M. V. (1999). Text compression as a test for artificial intelligence. In *Proceedings of the National Conference on Artificial Intelligence, AAAI* (pp. 970–970).

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In *Visual Information Processing, ed. W. Chase* (pp. 283–308). New York: Academic Press.

Oppy, G., & Dowe, D. L. (2011). The Turing Test. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Stanford University. Http://plato.stanford.edu/entries/turing-test/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainnment tests*. Danmarks paedagogiske Institut.

Raven, J. C., Court, J. H., & Raven, J. (1992). *Manual for Raven's Progressive Matrices and Vocabulary Scale*. San Antonio, TX: Psychological Corporation.

Sanghi, P., & Dowe, D. L. (2003). A computer program capable of passing I.Q. tests. In P. P. Slezak (Ed.), *Proc. of the Joint International Conference on Cognitive Science, 4th ICCS International Conference on Cognitive Science & 7th ASCS Australasian Society for Cognitive Science (ICCS/ASCS-2003)* (pp. 570–575). Sydney, NSW, Australia.

Schaie, K. W. (1985). *Schaie-Thurstone adult mental abilities test manual*. Consulting Psychologists Press.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *Autonomous Mental Development, IEEE Transactions on*, *2*, 230–247.

Shettleworth, S. (2010). *Cognition, evolution, and behavior*. Oxford University Press, USA.

Shettleworth, S. J. (1993). Where is the comparison in comparative cognition?: Alternative research programs. *Psychological Science*, (pp. 179–184).

Shettleworth, S. J. (2009). The evolution of comparative cognition: Is the snark still a boojum? *Behavioural processes*, *80*, 210–217.

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review*, *70*, 534.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis.*. Appleton-Century.

Solomonoff, R. J. (1962). Training sequences for mechanized induction. *Self-organizing systems, eds., M. Yovit, G. Jacobi, and G. Goldsteins*, *7*, 425–434.

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

Solomonoff, R. J. (1967). *Inductive Inference Research: Status, Spring 1967*. RTB 154, Rockford Research, Inc., 140 1/2 Mt. Auburn St., Cambridge, Mass. 02138, July 1967.

Solomonoff, R. J. (1978). Complexity-based induction systems: comparisons and convergence theorems. *Information Theory, IEEE Transactions on*, *24*, 422–432.

Solomonoff, R. J. (1985). The Time Scale of Artificial Intelligence: Reflections on Social Effects. *Human Systems Management*, *5*, 149–153.

Solomonoff, R. J. (1996). Does algorithmic probability solve the problem of induction? In D. L. Dowe, K. B. Korb, & J. J. Oliver (Eds.), *Proceedings of the Information, Statistics and Induction in Science (ISIS) Conference* (pp. 7–8). Melbourne, Australia: World Scientific.

Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *The American Journal of Psychology*, *15*, 201–92.

Sternberg, R. J. (2000). *Handbook of intelligence*. Cambridge University Press.

Strassmann, J. E., Klingler, C. J., Arévalo, E., Zacchi, F., Husain, A., Williams, J., Seppä, P., & Queller, D. C. (1997). Absence of within–colony kin discrimination in behavioural interactions of swarm–founding wasps. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *264*, 1565.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Syropoulos, A. (2008). *Hypercomputation: computing beyond the Church-Turing barrier*. Springer-Verlag New York Inc.

Thorndike, E. L. (1927). The law of effect. *The American J. of Psychology*, *39*, 212–222.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric monographs*, .

Tomasello, M., & Call, J. (1997). *Primate cognition*. Oxford University Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460.

Vernon, P. E. (1950). *The structure of human abilities.*. Methuen.

Wainer, H. (2000). *Computerized adaptive testing: a primer, 2nd Edition*. Lawrence Erlabaum Associate Publishers.

Wallace, C. S. (1990). Physically random generator. *Comput. Syst. Sci. Eng.*, *5*, 82–88.

Wallace, C. S. (2005). *Statistical and Inductive Inference by Minimum Message Length*. Springer-Verlag.

Wallace, C. S., & Boulton, D. M. (1968). An information measure for classification. *Computer Journal*, *11*, 185–194.

Wallace, C. S., & Dowe, D. L. (1999). Minimum message length and Kolmogorov complexity. *Computer Journal*, *42*, 270–283. Special issue on Kolmogorov complexity.

Watson, J. B. (1967). *Behavior: An introduction to comparative psychology*. A Henry Holt edition in psychology, 1967.

Webb, B. (2009). Animals versus animats: Or why not model the real iguana? *Adaptive Behavior*, *17*, 269–286.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised*. San Antonio, TX: Psychological Corporation.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*, 1–27.

Whiteson, S., Tanner, B., & White, A. (2010). The Reinforcement Learning Competitions. *The AI magazine*, *31*, 81–94.

Wilson, S. W. (1991). The animat path to AI. In J.-A. Meyer, & e. S. Wilson (Eds.), *From Animals to Animats* (pp. 15–21). MIT Press, Cambridge, MA, 2008.

Woergoetter, F., & Porr, B. (2008). Reinforcement learning. *Scholarpedia*, *3*, 1448.

Wolpert, D. H., & Macready, W. G. (2005). Coevolutionary free lunches. *Evolutionary Computation, IEEE Transactions on*, *9*, 721–735.

Zatuchna, Z., & Bagnall, A. (2009). Learning mazes with aliasing states: An LCS algorithm with associative perception. *Adaptive Behavior*, *17*, 28–57.