Dowe, DL.; Hernández Orallo, J. (2014). How universal can an intelligence test be?. Adaptive Behavior. 22(1):51-69. doi:10.1177/1059712313500502.

How universal can an intelligence test be?

David L. Dowe
Computer Science and Software Engineering
Clayton School of I.T., Bldg 63,
Monash University, Vic. 3800, Australia
david.dowe@infotech.monash.edu.au

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

Abstract

The notion of a *universal* intelligence test has been recently advocated as a means to assess humans, non-human animals and machines in an integrated, uniform way. While the main motivation has been the development of machine intelligence tests, the mere concept of a universal test has many implications in the way human intelligence tests are understood, and their relation to other tests in comparative psychology and animal cognition. From this diversity of subjects in the natural and artificial kingdoms, the very possibility of constructing a *universal* test is still controversial. In this paper we rephrase the question of whether universal intelligence tests are possible or not into the question of how universal intelligence tests can be, in terms of subjects, interfaces and resolutions. We discuss the feasibility and difficulty of universal tests depending on several levels according to what is taken for granted: the communication milieu, the resolution, the reward system or the agent itself. We argue that such tests must be highly adaptive, i.e., that tasks, resolution, rewards and communication have to be adapted according to how the evaluated agent is reacting and performing. Even so, the most general expression of a universal test may not be feasible (and, at best, might only be theoretically semi-computable). Nonetheless, in general, we can analyse the universality in terms of some traits that lead to several levels of universality and set the quest for universal tests as a progressive rather than absolute goal.

Keywords: intelligence; universal tests; test interface; space-time resolution; machine evaluation.

1 Introduction

One common definition of intelligence is given as the ability "to operate successfully in a wide variety of environments" (Russell and Norvig, 2005, chapter 2, p.32). This view implicitly acknowledges some kind of cognitive adaptation to new environments, instead of a view of success (or survival) in a narrow variety of environments, a niche, which can be obtained through evolution. Also, it is important to understand that the "wide variety of environments" must be considered in a cognitive rather than a physical way, as many environments would be just lethal to many (or all) life forms, independently of their intelligence. As a result, human intelligence is usually assessed by the behaviour in a variety of situations where different cognitive abilities and, ultimately, behaviours lead to different outcomes.

It is now commonly agreed that intelligence is not specific to humans. More and more life forms are attributed as also having some kinds or degrees of intelligence. The number and ranges

of species *claimed* to be in this category have been increasing in the past decades, from mammals to molluscs, from swarms to plants (Trewavas, 2005). Nonetheless, there is still an important debate as to whether some of them just exhibit a very complex behaviour (evolutionary species adaptation) rather than a truly intelligent behaviour (individual, non-hardwired adaptation). This already complicated picture of intelligent forms is (or will be) completed by machines and robots of many kinds, some of them already flaunting some kinds of intelligent behaviour. Similarly, there is a recurrent debate on whether the behaviour is authentically intelligent or just pre-programmed. And, finally, to really make the picture full, we should also consider hybrids (e.g., people augmented with pen and paper, with electronic devices and Internet connection, etc.) and communities (either homogeneous or heterogeneous), where the notions of intelligence and mind may diverge.

In this context, (Hernández-Orallo and Dowe, 2010) introduced the notion of 'universal (intelligence) test', as a test that could be administered to any kind of subject, at any speed, and interrupted anytime. The notion of a 'test for all' soon generated widespread interest (Kleiner, 2011; Biever, 2011), even though the proposal highlights some concerns about the difficulty (or even the feasibility) of such an idea, especially if we consider the evolution of comparative psychology in the past century. In fact, a new discipline called *universal psychometrics* (Hernández-Orallo et al., 2013) is suggested as an umbrella area of convergence for the general evaluation of cognitive abilities of any kind of subject (with or without universal tests).

While most of the above work has been motivated by the lack of proper tools to evaluate machine intelligence, the same rationale and principles could be used for natural subjects. In particular, in this work we analyse the notion of universal intelligence test in a more abstract way and analyse how universal an intelligence test can be. In particular, we will try to answer several questions: Can a test cover any kind of natural or computational subject, operating at any signal resolution and time rates? What resolution aspects should we consider? How crucial is the interface in a universal test? How can we dare measure the intelligence of subjects that we struggle to detect? Do we need intelligence to detect and measure some other intelligent subjects? How are validity, reliability and efficiency affected by a generalisation of a test?

In what follows we will go through these questions with a most general perspective, by considering that subjects can be machines, humans, non-human animals, plants, and other natural, artificial or hybrid systems. We start in section 2 with some necessary notions for the rest of the paper, such as the space of cognitive abilities. Section 3 follows with an overview of the wide range of intelligence tests in natural systems and machines. Section 4 describes what a universal test is and how it depends on the use of different interfaces for different kinds of subjects. Section 5 further generalises this idea by discussing possible configurations of time and resolution, and sketches a general procedure for an adaptive universal test. Section 6 examines more extreme cases where we do not know the rewarding system or we do not even recognise (in the beginning) the agent we want to evaluate. Finally, section 7 introduces a hierarchy of test universality levels according to several factors. We discuss the difficulty of universal tests according to this hierarchy and their feasibility for more constrained (but easier) applications that are expected to be common in the near future.

2 The space of cognitive abilities

As we mentioned in the introduction, we are concerned about the interactions between systems and environments in cognitive terms. Even if cognition is increasingly regarded as an *embodied* process with the rest of the physical characteristics of the system, it is still possible to distinguish

(with more or less difficulty) between *physical* and *mental* (or cognitive) processes. Cognitive systems, however, are complex systems that are usually difficult to understand completely. As a result, we describe them through the use of mental states and abilities. This is known as the ascription approach, which is not only applied to biological but also to artificial systems: "to ascribe certain beliefs, knowledge, free will, intention, consciousness, abilities or wants to a machine or computer program is legitimate when such an ascription expresses the same information about the machine that it expresses about a person" (McCarthy, 1979).

In this work we focus on cognitive *abilities* and their measurement. The use of the ascription approach (as humans have always done) raises doubts about an objective evaluation of a cognitive system: "The extent to which we regard something as behaving in an intelligent manner is determined as much by our own state of mind and training as by the properties of the object under consideration. [...] With the same object, therefore, it is possible that one man would consider it as intelligent and another would not;" (Turing, 1948, page 19).

Psychometrics tries to avoid this subjectivity objection by objectively defining how each ability has to be measured. Consequently, cognitive abilities (or mental abilities, according to Thurstone, 1938) are not only commonday linguistic terms (as with the ascription approach) but scientifically useful constructs to describe the behaviour of cognitive systems. As psychometrics has shown, psychological traits are high-order descriptions of a system that —if well-estimated, such as the weight of a physical object— are reliable, stable and predictive about many facets in life. Of course, cognitive abilities do not completely describe the subject or make it fully predictable, but determine what the subject is able to do, at least on average on a kind of *cognitive task*.

A cognitive task is "any task in which correct or appropriate processing of mental information is critical to successful performance" (Carroll, 1993, page 10). From here, "a cognitive ability [..] concerns some class of cognitive tasks, so defined". So, the "ability is defined in terms of being able to perform something", namely a class of cognitive tasks. From this association between abilities and classes of tasks, we see that by merging two cognitive task classes we get a more general cognitive task class, and a more general ability. Typically, this is studied in a hierarchical way, starting with the so-called elementary cognitive tasks (Carroll, 1993, page 11) (closely related to the notion of primary mental abilities of Thurstone, 1938).

Intelligence is usually understood as a cognitive ability. However, the interpretation of the set of tasks that correspond to intelligence is still the subject of research (and controversy). It is either seen as a very general ability that integrates many other abilities, or it is seen as a more specific one (e.g., the g-factor) that helps on many kinds of tasks but is not precisely optimal in most of them. Note that these two different views are consistent with the definition of intelligence as expected performance given in the introduction and the notion of cognitive ability.

One problem with precisely defining intelligence is that we usually associate it with phenomena such as mind and consciousness: "we seem to be attributing minds to the things we thus interpret" (Dennett, 1971). While the association is not usually made for some cognitive abilities (e.g., mindless computers excel in memory and calculation), mind (and consciousness) is usually seen as indissoluble with intelligence. However, as we broaden the spectrum of cognitive systems, we see that we can recognise intelligence in 'mindless' systems (e.g., a swarm). In the opposite sense, we can attribute a mind to some systems with very limited intelligence (e.g., people with certain strong neurological disorders). In fact, behind some intelligent behaviour there could be many minds, as happens with collectives. This is, in our opinion, one of the strongest objections to Searle's Chinese room (Searle, 1980) (apart from the compression objection, first introduced in Dowe and Hajek, 1997), a paradigmatic case of the confusion of conflating intelligence and mind. This also suggests ways of coping with a variety of cognitive abilities by including subsystems

(or modules, in Fodor's terms, 1983) that excel on each ability separately. However, in most cognitive systems (from swarms to brains), many individual components are not able to show any cognitive ability on their own, and the abilities only appear as an emergent property of an associated topology, protocol or coordination process.

Overall, the consideration of intelligence as a cognitive ability clarifies how measurement is performed, as a purely behavioural process. A cognitive test is simply any instrument to measure a cognitive ability, generally consisting in administering instances of the task class that define the ability. For instance, memory is an ability that is usually measured by a memory test, which is composed of a variety of memory tasks. Similarly, intelligence is measured by an intelligence test, which is composed of a variety of tasks, as in IQ tests.

3 Different subjects, different tests

Intelligence, in particular, and cognition, in more general terms, are nowadays associated with a diversity of species in the natural world. This diversity goes far beyond humans and the animal kingdom, as intelligence has also been recognised (or claimed) in swarms, plants, fungi, immune systems, bacteria, genomes and metabolic systems (Trewavas, 2005). The picture is completed by machine intelligence and extraterrestrial intelligence (Vakoch, 2011; Edmondson, 2012), collectives and hybrids of any of these systems. We will now give a short overview of the tests that are used to evaluate the cognitive abilities of all these kinds of systems.

Psychometrics (see, e.g., Sternberg, 2000; Borsboom, 2005) is the most mature (and now robust) discipline in terms of intelligence evaluation, but it is also the most anthropocentric one, as it focusses on human intelligence. As for the measurement of intelligence and other cognitive abilities, one remarkable characteristic in psychometrics is that we can have different tests for the same ability, depending on the kind of subject. For instance, we have different tests for children, disabled people, etc. It is important to note that while the tests (the instruments) are different, the ultimate purpose is to measure the same ability.

When looking at ways of evaluating cognitive abilities in a general way, comparative psychology (and cognition) (Shettleworth, 2010; Shettleworth et al., 2013) is the place to look. First, it has made a very important impact by looking at the problem in a less anthropocentric way. Second, it usually deals with the problem of using different tests for different species in order to measure the same ability. Third, it is an excellent source of how interfaces can be designed. Any species may be a subject of study for comparative psychology. However, most research has been devoted to animals with sophisticated cognitive processes, such as mammals (most especially apes, dogs and cetaceans), birds (most especially corvids) and some cephalopods. Also, the ease of experimentation has always been a factor, which explains the high number of studies for rats. Animal cognition usually refers to tasks and experiments (Shettleworth, 2010; Shettleworth et al., 2013), rather than 'tests', as in psychometrics. The tasks are used to detect and evaluate "basic processes", such as perception, attention, memory, associative learning and the discrimination of concepts, as well as more sophisticated physical or social abilities (Shettleworth et al., 2013).

As we are most interested in intelligence and related abilities, we have to pay attention to more recent work in comparative psychology that performs batteries of 'tests' with several individuals in order to use factor analysis and other mathematical tools that are common in psychometrics. As a very significant example, we succintly present a paradigmatic example of how "a comprehensive battery of cognitive tests [is applied] to large numbers of two of humans' closest primate relatives, chimpanzees and orangutans, as well as to 2.5-year-old human children" (Herrmann et al., 2007). While the goal of the study is to analyse some hypotheses about the

importance of social cognitive skills, the battery of tests includes both 'physical' tasks (such as space memory, working with quantities and causality) and 'social' ones (such as social learning, communication and theory of mind). In fact, the results show that chimpanzees are comparable to 2.5-year-old humans on the physical domain, while the difference appears in the social domain. One of the most significant features of this research is how the same task is presented in a different way to the three species. We will get back to this issue in Section 4.

At a presumably distant location on the spectrum of life forms, there are some organisms where some kind of minimal cognition (van Duijn et al., 2006) has been found, such as plants, fungi or bacteria (Trewavas, 2005). On other occasions, the cognition processes appear or are studied in a subsystem of the organism (immune systems, genome and metabolic systems, Trewavas, 2005) or as an aggregation of many organisms (swarms). These biological types of cognitive systems are sometimes referred to as non-neural organisms (Ginsburg and Jablonka, 2009), i.e., organisms without a neural system. While there seem to be genuine (minimal) cognitive processes in many of these systems, it is highly debatable whether we can use the term 'intelligence' for these systems. Some of the arguments in favour of plant intelligence are based on examples of classical conditioning (Haney, 1969) or some kinds of complex behaviour (Applewhite, 1975), even though some of these findings and claims have also been disputed (Sanberg, 1976). Trewavas, the most prominent advocator of plant intelligence, argues that plants are able to do basic problem solving and to predict the future (Trewavas, 2005). Examples include the prediction of future shade or the prediction of when water is to be provided. The question of whether complex regulatory (control) systems have to be considered intelligent or not (as in so-called intelligent control) is an ill-defined question, as it depends on where we put the threshold. A much more productive approach is to define traits we want to consider, and give measurements for different species and individuals, using different components of intelligence. This is exactly what he proposes with the use of intelligence 'rosettes', a kind of rudimentary factorial analysis, which is basically how psychometrics started. However, the traits Trewavas wants to measure are very *phytocentric*: "the most relevant are flexibility in leaf weight:area, speed of new leaf production, sensitivity to shade, flexible operation of photosynthesis, stomatal sensitivity (closing and opening speed after perturbation) and abscission sensitivity. Other traits need identifying and quantifying, and could then be included along with equivalent root and stem traits" (Trewavas, 2005). Any measurement using these traits is then not easily comparable to other traits used in comparative psychology (and even less for psychometrics). Nonetheless, the study of plant cognition reveals that reaction, adaptation and learning may take place at very different rates than happens with animals, that interaction with the world is mostly driven by light and chemical reactions (with very different mechanisms than vision and olfaction in animals) and that cognition may be distributed and originate in peripheral parts of the organism, such as roots.

If there is a strong debate about the intelligence of these non-neural cognition systems, the controversy is still stronger for the possibility and detection of machine intelligence. As a start, the terms *artificial intelligence* and *machine cognition* are frequently used nowadays to describe systems that lack intelligence and possibly any authentic (embodied) cognition. Consequently, instead of discussing whether these terms are appropriate or not, we will again analyse the root of the problem: the lack of agreement of what intelligence is and how it should be measured.

Over the past decades, there have been many proposals to evaluate AI artefacts or areas of the discipline. The Turing Test (Turing, 1950; Oppy and Dowe, 2011), the total Turing Test (Schweizer, 1998), sensorimotor variations (Neumann et al., 2009), the Bot Prize (Hingston, 2010), etc., do not really measure intelligence. The Turing Test (and variants) are, by definition, anthropomorphic, and evaluate *humanness*. Also, it presents some other problems, such as being non-gradual and non-factorial (see, e.g., Hernández-Orallo, 2000a, for a discussion).

Other approaches —such as the machine intelligence quotient, MIQ, (Zadeh, 1976; Bien et al., 2002) and CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) (von Ahn et al., 2004)— are defined in terms of what current AI technology is able to do today. For instance, CAPTCHAs are tasks that most humans can do easily while current technology cannot, such as reading distorted text. Figure 1 shows an example of a CAPTCHA. In fact, they are not used as *tests* of intelligence but just, as the acronym indicates, to tell humans and machines apart.

Figure 1: An example of a distorted string of characters ('smwm') that may serve as a CAPTCHA. There are also audio CAPTCHAs for the blind. (The image is in the public domain from http://commons.wikimedia.org/wiki/File:Captcha-smwm.svg)

There are also some approaches to measure machine performance based on competitions, such as the RL competition (Whiteson et al., 2010), Robocup (Kitano et al., 1997), the general game playing competition (Genesereth et al., 2005), the planning competition (Long and Fox, 2003), etc. These are very specific benchmarks, and they only evaluate a very specific ability for which contestants are designed in a highly specialised way. In other words, these are not general intelligence tests, they are just tests for very specific tasks and systems.

Of course, one can argue that different tests are required to measure different abilities (e.g., planning, visual recognition, natural language, etc.), but is it reasonable to have different tests for the same ability (e.g., intelligence or learning ability) if the subjects are different? As a response to this we should mention that there have been some advocates for the use of the same tests for (at least) humans and machines, namely the use of IQ tests for machines. The roots can be traced to the development and understanding of cognitive models, such as the works of Evans (1963; 1964; 1965) and, indirectly, those of Simon and Kotovsky (Simon and Kotovsky, 1963; Kotovsky and Simon, 1990). Nowadays, there is a field known as "psychometric AI" —not to be confused with universal psychometrics (Hernández-Orallo et al., 2013)—, where IQ tests are used to improve and evaluate artificial intelligence systems. Nonetheless, the most explicit vindication of the use of IQ tests for machines has been recently made by Detterman, editor of the *Intelligence* journal, as a response (Detterman, 2011) to specific domain tests and landmarks. In other words, this view claims that human-level (machine) intelligence should be measured with human-level intelligence tests, i.e., IQ tests.

However, there are many objections to the use of IQ tests for non-human agents, not only for animals (see, e.g, Shettleworth, 2010, sec. 1.1.4) but most especially for machines (see Dowe and Hernández-Orallo, 2012 for a full discussion). The main argument is that it is possible to find or construct non-intelligent agents that can score well on classical IQ tests, as has been demonstrated with very small programs (in 2003, Sanghi and Dowe implemented a small program in Perl which could score relatively well on many IQ tests). This raises serious doubts about the real implications of any non-human (natural or artificial) system that is evaluated using IQ tests (e.g., Eliasmith et al., 2012; Yong, 2012). It is not clear either that a generalisation or revision of current IQ tests to make them more general is a sound pathway, as we could end up with tests that work with a specific population or in terms of a current technology, but need to be updated recurrently as the population is enlarged and AI technology improves (as happens with the notion of Machine Intelligence Quotient, Zadeh, 1976; Bien et al., 2002, or, more blatantly, with CAPTCHAs, von Ahn et al., 2004).

Despite the variety of tests for humans, non-human biological systems and machines, there is a thing in common for all these tests: they lack a formal (and in most cases, principled) notion

$k = 9$	:	a, d, g, j, ...	Answer: m
$k = 12$	:	a, a, z, c, y, e, x, ...	Answer: g
$k = 14$	:	c, a, b, d, b, c, c, e, c, d, ...	Answer: d

Figure 2: Examples of series of Kt complexity 9, 12, and 14 used in the C-test (Hernández-Orallo, 2000a).

of intelligence from which the tests are derived. In other words, the ability is finally defined from what the test measures instead of deriving a test from a principled definition of the ability. At most, for the tests discussed so far, there may be a conceptual analysis of the nuts and bolts of the ability prior to devising the task(s) that may correspond to the ability. However, even in these cases, the link is usually made at a non-mathematical level. Also, the derivation of task difficulty is not obtained from a mathematical definition of the ability but as a refinement of the experimental results of a given population. In the end, this is the same way that thermometers and other measuring devices were designed in the past, without precisely knowing the exact definition of the magnitude to be measured and the physical processes involved.

As an alternative to these approaches, in the past two decades, there have been several efforts to derive definitions and tests of intelligence from computational principles. The relevance of (algorithmic) information theory (AIT), or Kolmogorov complexity, to this goal was first hinted at by Chaitin (Chaitin, 1982, sec. 6, f) before being independently elaborated upon in (Dowe and Hajek, 1997; Hernández-Orallo and Minaya-Collado, 1998; Mahoney, 1999; Hernández-Orallo, 2000a,c,b) with a series of tests noting the relevance and importance of the notions of (algorithmic) information theory (or Solomonoff-Kolmogorov complexity) and (two-part) compression (Solomonoff, 1964; Wallace and Boulton, 1968; Wallace and Dowe, 1999). An example of one of these tests formally derived from computational principles is shown in Figure 2, which resembles some exercises found in IQ tests but with a principled generation and assessment of difficulty.

Later on, (Dobrev, 2000, 2005) proposed a definition of (artificial) intelligence as an aggregate of performance in a wide range of environments, where the set of environments is described by Turing machines and bounded by Kolmogorov complexity. A similar approach was given by Legg and Hutter, 2005, 2007, who built upon re-inforcement learning by also using AIT, putting a Solomonoff-weighted prior distribution over single-agent environments. A measure (or definition, but not a test) of intelligence could be theoretically obtained (in the limit) by seeing what score the agent would obtain after infinite time in each of the infinitely many environments. There are several issues about the feasibility and exact interpretation of such a measure, as raised by (Hibbard, 2009; Hernández-Orallo and Dowe, 2010; Dobrev, 2013), among others.

4 The realisation of universal tests: interfaces

From the previous series of contributions on definitions, measures and tests based on AIT, the project anYnt[1] was set to analyse the possibility of constructing the first universal, formal, but at the same time practical, intelligence test. The term universal had been used and understood in different ways by many of the previous proposals, frequently in terms of concepts such as *universal* Turing machine, or the use of the term 'universal distribution' for any Solomonoff-weighted prior distribution. In this anYnt project, however, the term 'universal' has a more common meaning and refers to the applicability of the test to any kind of individual.

[1] http://users.dsic.upv.es/proy/anynt/

Pursuing this universality, taking into account the limitations of (Legg and Hutter, 2007) and other previous approaches, (Hernández-Orallo and Dowe, 2010) introduced an adaptive test which was *anytime* (able to be interrupted at anytime giving a more accurate result as more time is given) and also (supposedly) *universal*—i.e., applicable to machines, humans and non-human animals alike. While also based on algorithmic information theory, there are some distinctive features of this test. First, the class of environments is carefully selected to be discriminative. Second, while environments are randomly sampled from that class using an a priori distribution (starting with very simple environments), the complexity of the environments adapts to the subject's performance. Third, the test also considers time and, similarly, the speed of interaction adapts to the subject's performance. Finally, the result is not an average of results in all possible environments but an aggregation of how far an agent can reach in terms of the complexity and speed of the environment.

While some of these features are related, we are mostly concerned in this paper about the alleged universality of the test. If feasible, this universality should allow the comparison of very different subjects with the same tests. As a first proof-of-concept, (Insa-Cabrera et al., 2011) attempted an implementation of the test using the environment class introduced in (Hernández-Orallo, 2010). Actually, humans and AI agents (using some classical reinforcement learning algorithms) were evaluated using the same test. While the exercises were exactly the same, the interface was adapted to each kind of agent (we will show these interfaces in Figure 7). The general idea of a test for all was illustrated, but the results did not show a clear victory of humans over AI agents. There are several possible explanations for this (Hernández-Orallo et al., 2011; Hernández-Orallo et al., 2012; Insa-Cabrera et al., 2012): it was a prototype, it was not adaptive as per the original proposal (Hernández-Orallo and Dowe, 2010), there was no noise, patterns had low complexity, the environment class was quite limited, no social behaviour or other factors were evaluated, to name a few. In any case, the results do not indicate that the prototype is not a universal test, but rather that it does not measure intelligence properly.

In fact, the main feature of a universal test is that the task must be exactly the same while the interface has to be customised for each individual. In other words, a universal test is composed of a task and an infinite number of possible interfaces. Figure 3 shows this view of interfaces and tasks.

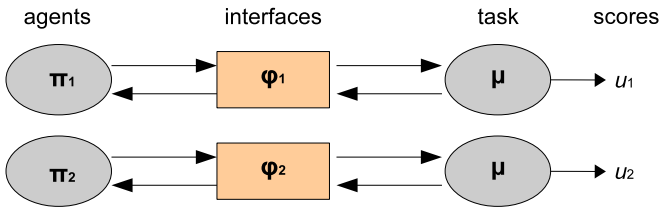


Figure 3: Two interfaces ϕ_1 and ϕ_2 for the same task μ in order to evaluate two different kinds of agents π_1 and π_2.

As the first of a series of examples, Figure 4 shows how a typical task found in IQ tests, such as Raven's Progressive matrix (Raven et al., 1992), can be presented with two different interfaces in order to evaluate sighted and blind people.

The application of a single task to different species is an everyday chore in comparative psychology, as is the choice of interfaces that are able to engage an animal to do a task (typically by the use of rewards) without adulterating the task itself. However, when the number of species for a single test is limited or just one, the task and the interface are sometimes blurred. Even when the species are akin, such as primates, the interface must be very carefully designed, in

Figure 4: Two different interfaces for a Raven's Progressive Matrix problem (Raven et al., 1992), a usual one and a version adapted for blind people.

terms of presentation of the problem, but also in terms of ensuring that the subjects are calm and motivated for the task. For instance, while the interfaces in Figure 5 for the experiments performed in (Herrmann et al., 2007) look fairly similar, the choice of rewards is highly important. Also, a familiar atmosphere was crucial: human children had to be evaluated on top of their mother's laps, and non-human apes had to be evaluated in an isolated but familiar context.

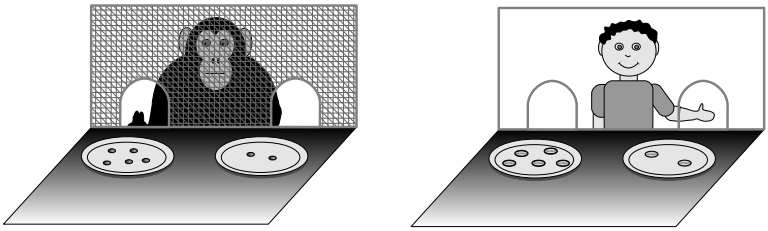


Figure 5: Two similar interfaces for the 'relative number task', as used for chimpanzees and small human children in (Herrmann et al., 2007).

In general, interfaces can be very specific for a species or particular for the task. Figure 6 (left) shows an original (an old) interface for number matching used for birds (Koehler, 1941). Figure 6 (right) shows a sophisticated interface for a maze task, which is just slightly particularised for two very different kinds of subjects: rats and robots (Barrera et al., 2011).

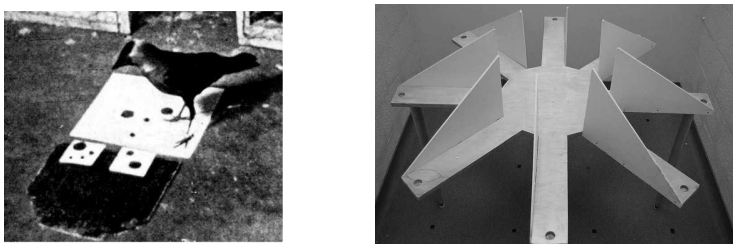


Figure 6: Left: an interface for a number matching task used for birds (Koehler, 1941). Right: an eight-arm maze, very similar to the one used for rats and robots in (Barrera et al., 2011) (Creative common licence from http://commons.wikimedia.org/wiki/File:Simple_Radial_Maze.JPG).

Finally, Figure 7 shows two different interfaces for the test developed in (Insa-Cabrera et al., 2011, 2012), one for humans and another one for a reinforcement learning agent such as Q-learning.

The meaning and emphasis of a universal (cognitive) test is the intention to apply the task

Figure 7: Two possible interfaces for the test in (Insa-Cabrera et al., 2011, 2012). Left: A very raw (computer) interface as a character string. Right: A snapshot of an exercise using a graphical interface for humans. The (inconclusive) experimental results in (Insa-Cabrera et al., 2011, 2012) were interpreted assuming that each interface was (almost) optimal for each kind of subject, but this might well not be the case.

to any possible agent: including humans, non-human animals, plants, communities, hybrids, ..., and most especially machines, including androids and animats (Webb, 2009). Even for those agents that clearly lack the ability (e.g., a flea solving sudokus), the intention is to figure out a way to measure the ability and give a score. This most general way forces us to place the concept of a test in terms of a general, abstract and *computational* task. We start with an information-processing ability, and define it properly as a set (or distribution) of tasks with a scoring system. Only after this, we may start thinking about interfaces to apply the task to any possible system. This computational approach is now also present in comparative and animal cognition, and particularly in animal evolutionary linguistics, constructing tasks about the kind of grammars (regular, context-free or contextual) several species are able to recognise (Gentner et al., 2006; Pullum and Rogers, 2006; Hauser et al., 2007; Evans and Levinson, 2009; Pullum and Scholz, 2009). From there, the evaluators figure out the best way of turning these tasks into tests by the use of appropriate interfaces.

This first understanding of what a universal test is and its possible application to any natural or artificial system also suggest that the question of whether universal tests are possible or not could be further refined into a more informative question —re which tasks can be measured with a universal test and which cannot. In principle, we may be able to construct universal tests for every information-processing task but, at the other extreme, it may also be the case that there are tasks for which no universal test can be constructed.

Before addressing this question (or the general question in the title of this paper), we need to look at the interfaces more closely. One reasonable criterion for a fair interface is that it cannot add new information or hide existing information about the task. For instance, if one interface discloses part of the solution while another does not, then the test would be adulterated. One possible mathematical way of defining an interface is by the notion of bijection, as done in (Hernández-Orallo et al., 2013, sec. 4.2). Nonetheless, there might be computational costs associated with the bijection. It is not always easy to see this in a mathematical way as the same problem may have many different representations, e.g., with diagrammatic representations or symbolic sequences in a formal or natural language (Larkin and Simon, 1987). This distinction was already made (Simon, 1978) when analysing the differences (and difficulty) between two different representations for the same problem. He distinguished between "informational equivalence", where "the transformation from one to the other entails no loss of information" (Simon, 1978) and "computational equivalence", where "the same information can be extracted with about the same amount of computation" (Simon, 1978). This is related to the notions of explicit and implicit information and the notion of information gain taking computational effort into account (Hernández-Orallo, 2000b).

However, while the concept of informational equivalence is necessary and useful, the appli-

cation of the notion of computational equivalence is not so clear. First, we need to clarify that there must be a clear distinction between the way the problem is represented and the problem itself. For instance, if a problem is the addition of numbers or the extrapolation of a sequence of numbers, as in some IQ tests, the test should not take into account the abilities of the subject to recognise numbers by symbols, sounds or by any other way. In other words, how well an agent is able to grasp the representation is not to be measured. In fact, we should try to use the most appropriate representation, independently of whether there is an important transformation between the representations for two different kinds of systems. There are cases where the transformation between two representations is not straightforward (even NP-hard) and one representation may be more suitable for a system while the other is more suitable for a second system. For instance, a unary representation of numbers may be more suitable for animals, while a decimal (positional) representation may be more suitable for a human who is used to this. These numeral systems (and many others, such as the sexagesimal system used in many cultures and still in ours when measuring time, or the Roman numerals) can be converted into the others in polynomial time (this means that they are computationally equivalent in Simon's terms, as "the same information can be extracted from each by the same amounts of computation, up to a factor of proportionality" Simon, 1978). But there are numeral systems where this transformation could be computationally more expensive. However, in general, if we have two subjects which are familiar with different numeral systems (convertable or not in linear time), we should use the one that fits each best (perhaps analogously to testing humans in their native languages).

In order to provide the representation which is most computationally appropriate for the agent (from those that are representationally equivalent), we need to know the subject well or use some degree of exploration, as a representation will only be optimal in terms of the easiness or "efficiency of search for information and in the explicitness of information" (Larkin and Simon, 1987). And this appropriateness must take place when the test starts and cannot be assumed to develop as the result of some other abilities of the agent. For instance, in some cognitive tests, awkward and tricky representations may be used, precisely because these tests may focus on detecting the ability of the agent to do good pattern recognition, visual transformations, etc. This also happens when tests take long periods of time. But if the test is not devised to measure this (such as an arithmetic test or a sequential inductive inference or prediction test), the representation should be as familiar and explicit as possible. For instance, if we use cups and peanuts for a chimpanzee, it is because we know that the chimpanzee is able to recognise cups and peanuts, and that they are part of its internal representation and easy to connect to its rewards. This highlights how critical the interface is. In fact, this *vulnerability* can be exploited in order to distinguish between species by using some kinds of tasks that are related to the way each species perceives the world, as CAPTCHAs do, rather than on their ability to process that perceived information.

Even under these constraints, there are still many possible interface choices for a given agent. For instance, in Figure 7 we saw two different interfaces for the test developed in (Insa-Cabrera et al., 2011, 2012). While the information is the same and the (computational) effort to transform the observation from one to the other is straightforward (informationally and computationally equivalent), humans would typically score much better with the interface on the right. Clearly, this search for the optimal interface for each species (or individual) can be applied to any natural or artificial agent. In artificial agents, exercises are usually presented in terms of data structures (arrays, sets, etc.) instead of visual interfaces, except for those tasks which are precisely evaluating visual recognition or related abilities. For natural agents, things are becoming more and more complicated. The interfaces used for some animal species are extremely elaborate and, on many occasions, different interfaces are used for different individuals, according to the under-

standing and knowledge of the subject's behaviour (e.g., in zoos, aquariums and natural parks, local curators are usually interviewed in order to decide which kind of interaction and rewards may be more appropriate for each individual). In the case of plants or other natural systems, the interfaces certainly seem quite unlike those previously used for animals. In fact, these interfaces are frequently the key to discovering abilities in these systems (Trewavas, 2005; Haney, 1969; Applewhite, 1975; Sanberg, 1976) that were considered non-existent only a few years ago. Humans are not an exception here either, as a great amount of imagination is needed to devise some tests for disabled people, especially for deaf-blind people, using tactile interfaces (Arnold and Heiron, 2002) or other approaches (Mar, 1996; Vernon et al., 1979; Rönnberg and Borg, 2001).

In general, the interface must pay attention to the sensorimotor characteristics of each agent, including other milieux, such as chemical sensors in animals and plants, and different kinds of data transformation for machines. Many failures in the past have been caused by important mismatches and misconceptions of how animals (and computers) should interact, with a strong anthropocentric bias in our interfaces.

So the grand question about the notion of a universal intelligence test is whether a test is able to evaluate a completely unknown agent. The general idea is that we can only assume some minimal information about the environment or *milieu* (either physical or virtual) where the agent is placed. From here, a real universal test should try to find the best interface (signals, time rate, resolution, rewards) which leads to the best score, without providing additional information about the task at hand. As a result, a totally universal, adaptive test requires a search in this vast area of time rates and resolutions. In the end, this is what animal cognition has done in the past decades in a manual (scientific-oriented) way. Is it possible to envisage such a process in an automated way, at least in some restricted contexts and for some abilities? In other words, if we are given an environment with some agents, is it possible to have a test that tries to find the interface to better evaluate the agents' abilities? This is what we explore next.

5 Time, resolution and universality: test adaptation

In comparative cognition, interfaces are usually associated with physical things: a cage with a small door, a peanut as reward, a set of cups, a touch screen, a light bulb, a set of ropes, etc. Thinking of a test that is able to adapt to all these possible physical configurations (and do this automatically) is currently far beyond reach. Instead of this, in this section, we will consider variations of the same physical (or virtual) 'milieu'. For instance, given a screen we can think about many possible resolutions and colours, given an audio signal in a range of frequencies we can consider all the possible variations there. In fact, many human tests are still administered with a sheet of paper, and many different interfaces are still possible with this 'rudimentary' milieu.

If we focus on cognitive abilities as information-processing tasks we can fix the milieu and examine the possible variations around it. With this restriction, any interface is in the end a pair of input and output communication channels (in terms of information theory) with a given bandwidth. If we consider discrete interfaces, we can describe this in terms of the input/output resolution and a refresh rate, which can apply to any possible milieu here —auditive, visual or other. For instance, considering time, if the task is the addition of two natural numbers less than 10 represented in a unary system, and we agree on the representation of the numbers in the output channel, there is still the question of how much time the numbers are going to be displayed and how much time the agent is going to be allowed to give an answer. If we fix these values we make the evaluation possible for some agents but this also excludes others. For

instance, plants are now claimed to do some kind of cognition (Calvo-Garzón and Keijzer, 2011), but their time-scale is much slower than those of animals.

The anytime test introduced in (Hernández-Orallo and Dowe, 2010) arguably addresses part of the issue of time adaptively by starting with a very fast interaction rate and slowing it down when the results from the agent are not good. The direction of the time change (from very fast to slower interaction) is reasonable as many agents would also react appropriately if the interaction is neither too fast nor too slow, and starting at fast interactions makes the adaptation feasible in finite time. While this is a first approach for making a universal test adaptive in time, there are more issues in terms of time-scale than those reflected by (Hernández-Orallo and Dowe, 2010). Also, other kinds of resolutions are not considered (Dowe, 2013, sec. 4.4).

If we take a closer look at time rate, we see at least two time frames that could be (adaptively) increased/decreased:

- Working time, which can be the time between questions and answers in a questionnaire-like test or the time other agents (e.g., predators) take to make actions or the time the environment keeps rewards available. This is in fact what (Hernández-Orallo and Dowe, 2010) adapts. This time typically includes the time the agent needs to act (or write an answer).
- Exposition time: the amount of time the agent gets to be given the information before it is removed. This time frame is usually neglected unless this is the goal of a study, e.g., rapid observation of input (or other kinds of related abilities, such as photographic memory). However, if the exposition time frame is not appropriate, the agent may completely overlook some important data of a task.

Let us use the term *time configuration* for the set of parameters for a given working and exposition time. Sometimes we want a cognitive ability to consider time, such as measuring 'reaction time' (McCarthy and Donchin, 1981). But if time is not part of the ability (e.g., we may want to know the ability to sort a series of numbers, without considering speed) then we (or the test) need to find the optimal time windows for working and exposition time in order to make the test feasible.

A related, but different, thing is resolution. Although we typically think in terms of spatial resolution, it is very useful for the discussion that follows to think of a case where we consider an audio signal, where resolutions appear on the same signal, using, e.g., frequency and amplitude. For instance, Figure 8 shows how a sound signal can carry many different types of information at several resolutions (Wilson et al., 2007; Madsen et al., 2012).

Not only may the resolution be too coarse or too detailed for the agent to see any relevant pattern, but it can take infinitely many representations. Note that the detection of the appropriate resolution is different from any pattern that the signal may carry. This distinction is important, even though both things (resolution and pattern) are usually closely intertwined. Nonetheless, while animals (including humans) usually have innate preprocessing systems that may be used to capture some resolutions (and ignore others) and see patterns in them, it is possible to disentangle one thing from the other. For instance, many artificial pattern (image, speech, etc.) recognition systems have preprocessing devices that render the information ready (e.g., a bitmap or spectrogram) for the analysis of patterns. The complexity of resolution and the appropriateness of representation is one of the major issues in artificial intelligence, pattern recognition and machine learning. This issue, however, has been neglected by most machine intelligence tests that we reviewed in section 3 or exploited for other purposes, such as CAPTCHAs. In fact, it is relevant to mention here that there are some recognition tasks (recognising distorted letters, as in CAPTCHAs) that do not correlate with some other (higher-level) abilities using those letters. In fact, the ability of recognising distorted letters is used as a CAPTCHA because machines are

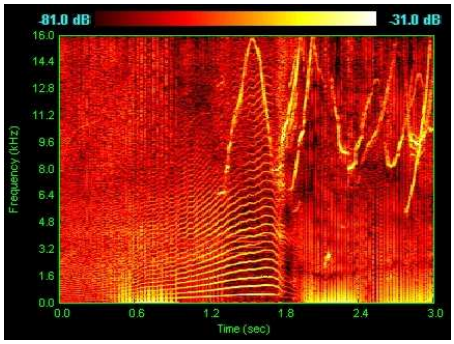


Figure 8: Dolphins perform a sophisticated use of the sound spectrum at different resolutions. Whistles and whines are used for communication and can be seen as horizontal striations and vertical lines, respectively. Clicks are used for echolocation, which are visible as strong, big wedges on the top-right of the spectrogram. (GNU-licensed image from wikimedia: http://commons.wikimedia.org/wiki/File:Dolphin1.jpg).

not able to do this well with current technology, not because distorted character recognition is a particular sign of intelligence.

Working with several resolutions and representations in order to find the most appropriate one may take important time overloads to process. This of course makes resolution and time also closely intertwined in real systems. From here, we can integrate both things by just defining a configuration θ as a pair of time configuration and resolution configuration.

So, if we want to evaluate a cognitive ability defined as a distribution of tasks M, i.e., a task class, we have to consider a set (or distribution) of configurations Θ. Then we can define:

$$U(\pi, M, \Theta) = \max_{\theta \in \Theta} \lim_{\tau \to \infty} \Upsilon(\pi, M, \theta, \tau) \tag{1}$$

where $\Upsilon(\pi, M, \theta, \tau)$ is any test on a family of tasks that is applicable to an agent π during time τ. For instance, we could do time-bounded adaptations of (Legg and Hutter, 2007), some of the non-adaptive versions introduced in (Hernández-Orallo and Dowe, 2010) or implemented in (Insa-Cabrera et al., 2011). Eq. (1) above defines (or generalises to) a universal test from a non-universal test. The expression $\Upsilon(\pi, M, \theta, \tau)$ is an aggregate over the set of tasks M. This must be based on the result $\Upsilon(\pi, \mu, \theta, \tau)$ on single tasks $\mu \in M$ and can be done in many different ways. One possibility is to weight by task probability: $\Upsilon(\pi, M, \theta, \tau) = \sum_{\mu \in M} \Upsilon(\pi, \mu, \theta, \tau) p(\mu)$ (as in Legg and Hutter, 2007), while another possibility is to define task difficulty and get the result in terms of this difficulty, as in (Hernández-Orallo and Minaya-Collado, 1998; Hernández-Orallo, 2000a; Hernández-Orallo et al., 2013).

This maximisation can be translated into the goal of finding the configuration such that the test result is optimal for the subject. This leads to *adaptive* tests, which search for the appropriate configuration, or at least one that gives an approximation of Eq. (1) above. In case we want the test to adapt, there is a need to have some interactive feedback from subject to testers, in terms of the score the subject is achieving. With this we are ready to introduce a first general procedure for an adaptive universal test. Figure 9 shows a general procedure for evaluating subject π in an adaptive way.

The relevant part of the previous procedure is how we select tasks and configurations, especially after the first iteration, using the history of results, tasks and configurations previously used. This

1. The test history H is initially empty.
2. Choose a task $\mu \in M$, a configuration $\theta \in \Theta$ and a time slot τ, taking H into account.
3. For a time τ, evaluate agent π with task μ at configuration θ, and store the result R.
4. Add $\langle \tau, \mu, \theta, R \rangle$ to the test history H.
5. Go to 2.

Figure 9: A test that adapts to time configuration and resolution configuration.

can be seen as an extension/generalisation of (Hernández-Orallo and Dowe, 2010), and ultimately of (Dobrev, 2005) and (Legg and Hutter, 2007) as well, with the appropriate modifications (Dowe, 2013, sec. 4.4). For instance, if the agent's score has been poor, we can either try to find a simpler task or change the configuration (which may imply a change in the time configuration, the resolution configuration or both). On the contrary, if the agent's score has been good, we would be tempted to keep the configuration and change to a more difficult (or more informative) task.

An important question is how to obtain the final result of the test. Ideally, if we were able to evaluate all possible configurations and all possible tasks for an infinite amount of time τ each, the result would be calculated by taking the aggregated performance on the task class (using its distribution or the difficulty function) for the configuration that has given the best result. In practice, in finite time, the test should start with configurations not taking too much time (so τ can be small) and try as many resolutions as possible as larger time slots are used.

So, a universal test for an unknown agent would be a test that makes all the possible efforts to find a configuration (best resolution and time configurations) such that the subject can be evaluated in optimal conditions. Not coincidentally, this is what animal cognition usually does when designing an experiment. In the most difficult cases, finding the correct configuration may take decades. Also, as we will discuss in section 7, we are never sure that the right configuration has even been found.

These difficulties appear even though we have already considered that the (physical) communication milieu is fixed (e.g., a sound channel or a screen) and we have also assumed that we recognise the agent and know its reward system. As mentioned above, we are not going to consider every possible physical milieu. Instead, in what follows we will consider an environment including both the evaluator and the evaluee (e.g., the real world, but most especially, virtual words, such as games, social networks, etc., because of the possible applications and the higher feasibility of a test of this kind).

6 Making contact: detecting intelligence

As said above, we now consider that evaluator and evaluee are agents in an environment. We assume that both can have an effect on the environment through actions and they can perceive changes in the environment. This does not necessarily mean that the environment has a spatial configuration where they can 'move', as in (Hernández-Orallo, 2010), but just that the possibility of interaction exists (e.g., an environment could just be a series of communication channels as in the matching pennies game, the Turing Test or any configuration in between, Hernández-Orallo et al., 2012). We also consider that the evaluee has a goal or reward system that is determined by the things that happen in the environment. This setting is assumed in multi-agent systems, games and many other virtual, robotic or physical environments, from artificial intelligence to

biology.

A first situation we can consider is when we know the evaluee's reward system so that the evaluator can directly use this by giving more or less rewards to the evaluee according to its performance. This sets an important advantage because we can condition the system to do things. Still, we do not know what communication channel or milieu to use, just how to administer rewards. So the big challenge here is to determine the way to communicate. To do this, the evaluator needs to perform actions to probe some reactions from the evaluee, in order to understand how to 'condition' the evaluee (to try) to perform a task. With animals, we typically use food for this (to approach them, make them trust the evaluator, etc.).

Even in a simplified virtual environment, as found in multi-agent systems or multiplayer games, recognising which channel the agents may use is not always easy. In general, there may be many possible channels, and communication can switch to more efficient communication channels with time. For instance, it is common that pre-established communication channels or protocols are omitted and agents end up communicating through other means that were originally not meant for communication (because the new channel is more efficient, can be concealed from other agents, or other reasons). The communication channel may be extremely original, as we can see in many examples of intra- or inter-species communication, from primates (Zuberbühler, 2000) to bacteria (Federle et al., 2003). Nonetheless, there are many kinds of communication which are not usually recognised as such explicitly, as the communication that takes place whenever predators and preys meet, where the very movements and peer positions are an authentic exchange of information. In biology, there is a huge variety of ways of contact and communication, and many are still to be unveiled. In virtual or technological environments this diversity is also very large now.

Performing a universal test under these circumstances would imply a generalisation of the procedure in Figure 9 by including a variety of channels (along with the diversity of configurations and tasks). In order to determine that the channel is working we would still rely on the reward system and start with very easy tasks, in order to raise the probability that the evaluee makes a correct action (possibly by chance) so it can get positive rewards and then start 'focussing' on the task.

Still, a much more challenging problem is when we do not know the agent's reward system. Without knowing the agent's goals (or reward system) we may detect reaction, but it is much more difficult to properly apply a test. The most natural option is to try to learn the agent's goals or rewards by observing habits, resistance to change, interacting, etc. This is again what ethologists can do, or what children do when playing with bugs or other animals in order to understand their behaviour. In order to automate this process (at least partially), there are several information-theoretic options that could be used to detect (and spur) agent-environment or agent-agent interaction (Tarapore et al., 2006; Williams and Beer, 2010), but other approaches exist (relational dynamics, information structure, and many others). Independently of what technique is finally used (and whether they can work in general or not), if the evaluator is able to discover the agent's goals then the test could start as in the previous case, using a generalisation of the procedure in Figure 9. In this case, some information about the interaction that was used to determine the rewards could also be reused as history for the selection of milieu, time configuration and resolution configuration.

Finally, the most challenging situation is when we have not even recognised the agent we want to evaluate. In this case, we are left in an environment (physical, robotic or virtual) and we need to discover whether there are intelligent forms there and try to evaluate them. One of the first caveats in this situation is that there may be many agents, and they can work collectively. As a consequence, even if we are able to recognise the individuals, we may fail at properly recognising

intelligence if this only appears as an emergent property of the collection of agents and not of the individuals (as in swarm intelligence).

Either in the form of an individual or a collective, this extremely challenging case has been occasionally discussed in biology but most especially in astro-biology. While astro-biology looks for physical signs of complex biological molecules, the detection of extraterrestrial intelligence through signals from outer space has been subject of much interest in the past decades (Vakoch, 2011) and has led to incipient fields such as 'astrocognition' (Dunér, 2011)(Vakoch, 2011, ch. 31). However, there is no clear methodology about how to do it (what to scan, as with the SETI project, and what to send, as with the Voyager space probes). Interestingly, there is a recent approach looking for "a set of universals", some "general cognitive principles" (Edmondson, 2012), that could be applicable to any system. These are seen as a necessity as, "without some informed analysis of what it might mean to be intelligent —yet radically different in terms of evolution, culture, biology, planetary location, and the like— we should not even begin to search for extraterrestrial intelligence" (Edmondson, 2012).

One problem is to recognise intelligence from an extraterrestrial transmission, and a related secondary problem is to be able to discern and interpret the message's intended communication. The interpreting of such a message would presumably be best done by the (Bayesian) algorithmic information-theoretic approaches of Solomonoff prediction (Solomonoff, 1964) and/or Minimum Message Length (MML) (Wallace and Boulton, 1968; Wallace and Dowe, 1999; Wallace, 2005). If the (algorithmic) information content (or Solomonoff-Kolmogorov complexity) of the communication is very low, this suggests something ultra-regular like a pulsar, analogous to a human repeating the same sound. If the Kolmogorov complexity is very high, this suggests unstructured random background gibberish, perhaps analogous to an incoherent human infant. For the message to stand out, it must be like, e.g., (human) language and have some structure without being totally repetitive. This should mean having some incrementality in its complexity, with sections of the message depending on previous sections. In terms of what to put in such a message or how to decode such a message, Wallace (private communication) (Dowe, 2008, sec. 0.2.5) considered explaining arithmetic, then eventually Turing machines, then eventually the Lyman series and Hydrogen, etc. Solomonoff likewise wrote much about training sequences (Solomonoff, 1962, 1984, 2010)(Solomonoff, 1967, sec. 7, dolphin talk).

This 'contact' problem corresponds very neatly to our extreme case of a universal intelligence test where a priori we do not even know about the existence of intelligence forms in an environment and, if they exist, where they are and what they look like. One important difference, though, is that in our case we assume the possibility of interaction, which (by modern technology) can only take place with interstellar communication with a very slow time-frame. This means that the principles could be the same, but the degree of repetition in the signals could be highly reduced when the interaction starts. When trying to systematise this, we could use yet another generalisation of the procedure in Figure 9 where the iteration would start looking for agents. The use of information-theoretic tools seems to be appropriate here as well. The mere recognition of an agent in these terms is ambitious as an agent can emerge (e.g., by autopoiesis) from very simple rules. For instance, these embodied, minimal cognitive agents have even been found (such as gliders) in very minimalistic environments such as Conway's game of life (Beer, 2004). As there may be many different kinds of agents we may need tools to determine those that are merely interactive, cognitive or finally intelligent. Any procedure or test that could be able to eliminate those agents that are not good candidates for further levels would be useful. For instance, once an interactive system is found we could first determine whether we have a cognitive system, adapting some of the tasks of the cognitive decathlon and related approaches (Anderson and Lebiere, 2003; Mueller and Minnery, 2008; Mueller et al., 2007; Mueller, 2008;

LEVEL	Configuration	Milieu	Rewards	Agent
1	✓	✓	✓	✓
2	×	✓	✓	✓
3	×	×	✓	✓
4	×	×	×	✓
5	×	×	×	×

Table 1: Several levels of universality of a cognitive test depending on the information we have about the time and resolution configuration, the communication milieu, the reward system and the agent itself.

Simpson Jr and Twardy, 2008; Chadderdon, 2008; Langley, 2011; Calvo-Garzón, 2003; Mueller and Minnery, 2008), —although many of these approaches are focussed on evaluating cognitive architectures rather than evaluating actual cognitive systems through multi-factorial scores.

7 Discussion

Jacob Bronowski, in one of his TV episodes of the 1973 BBC documentary "The Ascent of Man", asked the question about whether or not we would necessarily recognise intelligence if it were right beside us. Our answer to this question is that the difficulty of recognising (and ultimately) measuring intelligence would depend on a number of factors, as shown in Table 1. This table shows a gradual view which consists of five possible levels based on four criteria: knowledge of the configuration, milieu, rewards and agents. On occasions we may have some other combinations of these factors or we may have partial information about the configuration (or no information at all about the configuration but still some information about how the reward system works). Also, more information might be available about some agents than others. This suggests the use of Table 1 as an indication of the factors that need to be taken into account.

All of the cognitive test approaches that have tried to go beyond one type of system would fall at level 1, except perhaps the anytime universal test introduced in (Hernández-Orallo and Dowe, 2010). Consequently, Table 1 is not very informative as to whether there has been some progress in the past decades in the direction of more universal intelligence tests. In order to better show the current situation, Table 2 shows some cases of 'interspecies' cognitive tests, the kind of systems that have been evaluated and which kind of test they are. Note that for all the cases in Table 2 there is good knowledge of the kind of systems that are evaluated, in such a way that interfaces are designed very thoroughly, as we saw with examples in section 4. The less knowledge we have, the more difficult the evaluation will be, and we will get lower reliability, efficiency and (probably) validity.

As the evaluation becomes more difficult, the adaptability of the test becomes crucial. The test can be conceived in a passive (such as classical paper and pencil IQ tests), interactive (such as games, computerised adaptive testing, the anytime test, Hernández-Orallo and Dowe, 2010, or any generalisation of Figure 9) or ultimately intelligent way. In this regard, the notion of an 'intelligent intelligence test' may sound strange but it is the way intelligence has been detected and evaluated for many centuries, as in interviews and other personal (psychological) assessments. In fact, the Turing test is one example of an 'intelligent intelligence test', since it requires one intelligent subject to evaluate a possibly intelligent subject against a third intelligent subject.

It can be argued that the more intelligent the evaluator is the more effective the test can be. While this may be true (especially because a superintelligent being may be able to learn

Test	Abilities	Humans	Animals	Machines
Comparative Psychology	Usually narrow	✓	✓	×
(Herrmann et al., 2007)	Physical/Social range	✓	✓	×
IQ tests	Intelligence	✓	✓*	✓*
Turing Tests	Humanness	✓	×	✓
CAPTCHAs	Humanness	✓*	×	✓*
(Barrera et al., 2011)	Specific	×	✓	✓
BotPrize	Believability	✓*	×	✓*
(Insa-Cabrera et al., 2011)	Intelligence	✓*	×	✓*

Table 2: A non-exhaustive list of tests applied to more than one kind of subject. We mark with (*) those applications which do not really capture the ability that was originally conceived, or where there are doubts that the ability is properly evaluated. Hybrids, collectives and extraterrestrials are excluded from the table, as no test has been applied to these cases for more than one of the three kinds, except for some occasional tasks with swarms also applied to people or other animals.

the best procedures to do intelligence testing as they are discovered), it raises many questions about reliability and validity, since an intelligent being may not follow a clear procedure or may not even be fair and objective in the assessment. As there is a subtle line between adaptability and intelligence, we prefer to envisage universal intelligence tests which are highly adaptive but follow a formal and standardised procedure.

In terms of reliability, we can also see (from the way the procedure in Figure 9 works) that even when the proper configuration, milieu, reward system and agent are found then the assessment will still be an *under-estimation*. This originates from the way the overall result of the test is understood: the procedure tries to find the best conditions for administering the test (i.e., the measure is a maximum). As a result, the evaluation is biassed to under-estimate intelligence (as has happened for many years for non-human intelligence and has happened with other human cultures, and may happen with machines as well, Solomonoff, 1967, sec. 6, "The Problem of the Ambitious Subordinate"). This is a well-known problem in animal cognition (and also happens with other places where intelligence is sought, such as plants, bacteria, SETI, etc.). Moreover, this under-estimation also happens in human psychometrics, where the term "potential intelligence" is applied to "test potential" (Mahrer, 1958; Thorp and Mahrer, 1959; Little and Bailey, 1972), which is not to be confused with the term "potential intelligence" applied to the ability of becoming intelligent (Hernández-Orallo and Dowe, 2013).

From a computational point of view, we can say that evaluating intelligence (in a universal context) is, at most, semi-computable (i.e., it could be approximated by a computable function from below). This comes from the fact that given an environment that may contain intelligence, there are infinitely many configurations, milieux, reward systems or agents, and there may always be some of them that have not been explored (because of the halting problem, Turing, 1950). In theory, we could turn this into a computable function or even a function that can be calculated in bounded space and time, by assuming that configurations, milieux, reward systems or agents are resource-bounded.

Finally, there is a risk of over-estimation if the evaluator ends up training the agent instead of evaluating it. During a very long exploration for optimal configurations we may (inadvertently) train the agent we want to evaluate and make it more intelligent than it was. In fact, there are training sequences (Solomonoff, 1962, 1984, 2010)(Solomonoff, 1967, sec. 7, dolphin talk) for any universal Turing machine such that the machine becomes intelligent (in fact, as intelligent as we

want, as it can simulate any behaviour, as shown and fully discussed in Hernández-Orallo and Dowe, 2013). So there is, in principle, a problem in making intelligence tests too long, as the agents can be domesticated (trained) to become intelligent, unless the agents can be reinitialised for each iteration of the adaptive procedure. If the agent is trained (or just domesticated), the test mixes actual and potential intelligence (in the sense of Hernández-Orallo and Dowe, 2013), and the reliability and validity of the test are highly compromised.

Let us now summarise the results of our exploration about the limits and caveats of universal intelligence tests. We have seen how intricate the notion of universal intelligence test is and how difficult its implementation can be, as far as the challenge is set at the highest level in the hierarchy shown in Table 1. Nonetheless, the difficulty also depends on the environments and kinds of agents we want to explore. If we consider all possible natural and artificial agents in our physical world, a fully universal test working in a reasonable amount of time is just a theoretical idea. However, if we consider some simple environments (e.g., cellular automata, multi-agent environments or hybrids) or situations where we have part of the information of the hierarchy shown in Table 1, we may still establish universal intelligence tests for those niches.

In the end, the problem exists that there are current situations where we need to determine the intelligence of a completely unknown agent. This is becoming more and more common in virtual environments, such as social networks, games, working groups, member authorisation procedures, etc., where we may find agents whose intelligence is completely unknown. As discussed in the previous sections, the Turing Test, custom IQ tests and CAPTCHAs are unsatisfactory in general terms (and will be less so as artificial intelligence advances). As new artificial agents, living organisms, collectives and hybrids thereof become mainstream, we will need to better understand the concept of test universality and devise more effective universal tests for intelligence and other cognitive abilities.

Acknowledgements

We thank Paco Calvo for interesting discussions on plant intelligence and other kinds of non-neural intelligence, and for providing us with very useful pointers on this topic. Ultimately, this is an outcome of the encouraging atmosphere created by Manuel G. Bedia before, during and after the II ReteCog Meeting in early 2013. We also thank the editor and reviewer(s) for their comments, which have helped to improve this paper significantly and to make it more accessible to a broader audience. This work was supported by the MEC/MINECO projects CONSOLIDER-INGENIO CSD2007-00022 and TIN 2010-21062-C02-02, GVA project PROMETEO/2008/051, the COST - European Cooperation in the field of Scientific and Technical Research IC0801 AT.

References

Anderson, J. R. and Lebiere, C. (2003). The Newell test for a theory of cognition. *Behavioral and Brain Sciences*, 26(5):587–601.

Applewhite, P. (1975). Learning in bacteria, fungi, and plants. *Invertebrate learning*, 3:179–186.

Arnold, P. and Heiron, K. (2002). Tactile memory of deaf-blind adults on four tasks. *Scandinavian Journal of Psychology*, 43(1):73–79.

Barrera, A., Cáceres, A., Weitzenfeld, A., and 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.

Beer, R. D. (2004). Autopoiesis and cognition in the game of life. *Artificial Life*, 10(3):309–326.

Bien, Z., Bang, W. C., Kim, D. Y., and Han, J. S. (2002). Machine intelligence quotient: its measurements and applications. *Fuzzy sets and systems*, 127(1):3–16.

Biever, C. (2011). Ultimate IQ: one test to rule them all. *New Scientist*, 211(2829):42–45.

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

Calvo-Garzón, F. (2003). Nonclassical connectionism should enter the decathlon. *Behavioral and Brain Sciences*, 26(05):603–604.

Calvo-Garzón, P. and Keijzer, F. (2011). Plants: adaptive behavior, root-brains, and minimal cognition. *Adaptive Behavior*, 19(3):155–171.

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

Chadderdon, G. L. (2008). Assessing machine volition: an ordinal scale for rating artificial and natural systems. *Adaptive Behavior*, 16(4):246–263.

Chaitin, G. J. (1982). Godel's theorem and information. *International Journal of Theoretical Physics*, 21(12):941–954.

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, 68(4):87–106.

Detterman, D. K. (2011). A challenge to Watson. *Intelligence*, 39(2-3):77 – 78.

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(3):277–285.

Dobrev, D. (2013). Comparison between the two definitions of AI. *arXiv preprint arXiv:1302.0216*.

Dowe, D. L. (2008). Foreword re C. S. Wallace. *Computer Journal*, 51(5):523 – 560. Christopher Stewart WALLACE (1933-2004) memorial special issue.

Dowe, D. L. (2013). Introduction to Ray Solomonoff 85th Memorial Conference. In *Proceedings of Solomonoff 85th memorial conference*, volume LNAI/LNCS 7070, pages 1–36. Melbourne, Australia.

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

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

Dunér, D. (2011). Astrocognition: Prolegomena to a future cognitive history of exploration. In *Humans in Outer Space-Interdisciplinary Perspectives*, pages 117–140. Springer.

Edmondson, W. (2012). The intelligence in ETI - What can we know? *Acta Astronautica*, 78:37–42.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, C., and Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, 338(6111):1202–1205.

Evans, N. and Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(05):429–448.

Evans, T. (1963). *A heuristic program of solving geometric analogy problems*. PhD thesis, Mass. Inst. Tech., Cambridge, Mass., U.S.A. Also available from AF Cambridge Research Lab, Hanscom AFB, Bedford, Mass., U.S.A.: Data Sciences Lab, Phys and Math Sci Res Paper 64, Project 4641.

Evans, T. (1965). A heuristic program to solve geometric-analogy problems. In *Proc. SJCC*, volume 25, pages 327–339. vol. 25.

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.

Federle, M. J., Bassler, B. L., et al. (2003). Interspecies communication in bacteria. *Journal of Clinical Investigation*, 112(9):1291–1299.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psycholgy*. The MIT Press.

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

Gentner, T. Q., Fenn, K. M., Margoliash, D., and Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature*, 440(7088):1204–1207.

Ginsburg, S. and Jablonka, E. (2009). Epigenetic learning in non-neural organisms. *Journal of biosciences*, 34(4):633–646.

Haney, R. E. (1969). Classical conditioning of a plant: Mimosa pudica. *J. Biol. Psychol*, 11:5–12.

Hauser, M. D., Barner, D., and O'Donnell, T. (2007). Evolutionary linguistics: A new look at an old landscape. *Language Learning and Development*, 3(2):101–132.

Hernández-Orallo, J. (2000a). Beyond the Turing Test. *J. Logic, Language & Information*, 9(4):447–466.

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

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

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

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

Hernández-Orallo, J. and Dowe, D. L. (2013). On potential cognitive abilities in the machine kingdom. *Minds and Machines*, 23(2):179–210.

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

Hernández-Orallo, J., Dowe, D. L., and Hernández-Lloreda, M. V. (2013). Universal psychometrics: Measuring cognitive abilities in the machine kingdom. *Cognitive Systems Research, DOI:10.1016/j.cogsys.2013.06.001.*

Hernández-Orallo, J., Insa, J., Dowe, D. L., and Hibbard, B. (2012). Turing Tests with Turing machines. In Voronkov, A., editor, *The Alan Turing Centenary Conference*, volume 10, pages 140–156. EPiC Series.

Hernández-Orallo, J. and 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)*, pages 146–163. ICSC Press.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., and Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, Vol 317(5843):1360–1366.

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

Hingston, P. (2010). A new design for a Turing Test for bots. In *Computational Intelligence and Games (CIG), 2010 IEEE Symposium on*, pages 345–350. IEEE.

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

Insa-Cabrera, J., Hernández-Orallo, J., Dowe, D., España, S., and Hernández-Lloreda, M. (2012). The anYnt project intelligence test : Lambda - one. In Muller, V. and Ayesh, A., editors, *AISB/IACAP 2012 Symposium "Revisiting Turing and his Test"*, pages 20–27. The Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., and Osawa, E. (1997). Robocup: The robot world cup initiative. In *Proc. of the 1st international conference on autonomous agents*, pages 340–347. ACM.

Kleiner, K. (2011). Who are you calling bird-brained? An attempt is being made to devise a universal intelligence test. *The Economist*, 398(8723):82.

Koehler, O. (1941). Vom erlernen unbenannter anzahlen bei vögeln. *Naturwissenschaften*, 29(14):201–218.

Kotovsky, K. and Simon, H. A. (1990). What makes some problems really hard: Explorations in the problem space of difficulty. *Cognitive Psychology*, 22(2):143–183.

Langley, P. (2011). Artificial intelligence and cognitive systems. *AISB Quarterly*.

Larkin, J. H. and Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive science*, 11(1):65–100.

Legg, S. and Hutter, M. (2005). A universal measure of intelligence for artificial agents. In *Intl Joint Conf on Artificial Intelligence, IJCAI*, volume 19, pages 1509–1510.

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

Little, V. L. and Bailey, K. G. (1972). Potential intelligence or intelligence test potential? a question of empirical validity. *Journal of Consulting and Clinical Psychology*, 39(1):168.

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

Madsen, P. T., Jensen, F. H., Carder, D., and Ridgway, S. (2012). Dolphin whistles: a functional misnomer revealed by heliox breathing. *Biology letters*, 8(2):211–213.

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

Mahrer, A. R. (1958). Potential intelligence: a learning theory approach to description and clinical implication. *The Journal of General Psychology*, 59(1):59–71.

Mar, H. (1996). Psychological evaluation of children who are deaf-blind: An overview with recommendations for practice. *DB-LINK*.

McCarthy, G. and Donchin, E. (1981). A metric for thought: A comparison of p300 latency and reaction time. *Science*, 211(4477):77–80.

McCarthy, J. (1979). Ascribing mental qualities to machines. *http://www-formal.stanford.edu/jmc/ascribing/ascribing.html*.

Mueller, S. T. (2008). Is the Turing Test still relevant? A plan for developing the cognitive decathlon to test intelligent embodied behavior. In *19th Midwest Artificial Intelligence and Cognitive Science Conference, MAICS*.

Mueller, S. T., Jones, M., Minnery, B. S., and Hiland, J. M. H. (2007). The BICA cognitive decathlon: A test suite for biologically-inspired cognitive agents. In *Proceedings of Behavior Representation in Modeling and Simulation Conference, Norfolk*.

Mueller, S. T. and Minnery, B. S. (2008). Adapting the Turing test for embodied neurocognitive evaluation of biologically-inspired cognitive agents. In *Proc. 2008 AAAI Fall Symposium on Biologically Inspired Cognitive Architectures*.

Neumann, F., Reichenberger, A., and Ziegler, M. (2009). Variations of the Turing Test in the age of internet and virtual reality. In *KI 2009: Advances in Artificial Intelligence*, pages 355–362. Springer.

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

Pullum, G. K. and Rogers, J. (2006). Animal pattern-learning experiments: Some mathematical background. *Ms. Radcliffe Institute for Advanced Study/Harvard University*.

Pullum, G. K. and Scholz, B. C. (2009). For universals (but not finite-state learning) visit the zoo. *Behavioral and Brain Sciences*, 32(05):466–467.

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

Rönnberg, J. and Borg, E. (2001). A review and evaluation of research on the deaf-blind from perceptual, communicative, social and rehabilitative perspectives. *Scandinavian Audiology*, 30(2):67–77.

Russell, S. J. and Norvig, P. (2005). *Artificial intelligence: a modern approach, 1st edition*. Prentice Hall.

Sanberg, P. R. (1976). Neural capacity in Mimosa pudica: a review. *Behavioral biology*, 17(4):435–452.

Sanghi, P. and Dowe, D. L. (2003). A computer program capable of passing I.Q. tests. In Slezak, P. P., editor, *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)*, pages 570–575, Sydney, NSW, Australia.

Schweizer, P. (1998). The truly total Turing Test. *Minds and Machines*, 8(2):263–272.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457.

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

Shettleworth, S. J., Bloom, P., and Nadel, L. (2013). *Fundamentals of Comparative Cognition*. Oxford University Press.

Simon, H. A. (1978). On the forms of mental representation. *Perception and cognition: Issues in the foundations of psychology*, 9:3–18.

Simon, H. A. and Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review*, 70(6):534.

Simpson Jr, R. and Twardy, C. (2008). Refining the cognitive decathlon. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, pages 124–131. ACM.

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

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and control*, 7(1):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. (1984). Perfect training sequences and the costs of corruption - a progress report on induction inference research. *Oxbridge Research*.

Solomonoff, R. J. (2010). Algorithmic probability, heuristic programming and AGI. In *Proc. 3rd Conf. on Artificial General Intelligence. Advances in Intelligent Systems Research*, volume 10, pages 151–157.

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

Tarapore, D., Lungarella, M., and Gómez, G. (2006). Quantifying patterns of agent–environment interaction. *Robotics and Autonomous Systems*, 54(2):150–158.

Thorp, T. R. and Mahrer, A. R. (1959). Predicting potential intelligence. *Journal of Clinical Psychology*, 15(3):286–288.

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

Trewavas, A. (2005). Plant intelligence. *Naturwissenschaften*, 92(9):401–413.

Turing, A. (1948). Intelligent machinery. Report for national physical laboratory. Reprinted in Ince, D.C. (editor). 1992. Mechanical intelligence: Collected works of AM Turing.

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

Vakoch, D. A. (2011). *Communication with Extraterrestrial Intelligence*. S.U.N.Y. Press.

van Duijn, M., Keijzer, F., and Franken, D. (2006). Principles of minimal cognition: Casting cognition as sensorimotor coordination. *Adaptive Behavior*, 14(2):157–170.

Vernon, M. et al. (1979). Psychological evaluation and testing of children who are deaf-blind. *School Psychology Digest*, 8(3):291–95.

von Ahn, L., Blum, M., and Langford, J. (2004). Telling humans and computers apart automatically. *Communications of the ACM*, 47(2):56–60.

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

Wallace, C. S. and Boulton, D. M. (1968). An information measure for classification. *Computer Journal*, 11(2):185–194.

Wallace, C. S. and Dowe, D. L. (1999). Minimum message length and Kolmogorov complexity. *Computer Journal*, 42(4):270–283. Special issue on Kolmogorov complexity.

Webb, B. (2009). Animals versus animats: Or why not model the real iguana? *Adaptive Behavior*, 17(4):269–286.

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

Williams, P. L. and Beer, R. D. (2010). Information dynamics of evolved agents. In *From Animals to Animats 11*, pages 38–49. Springer.

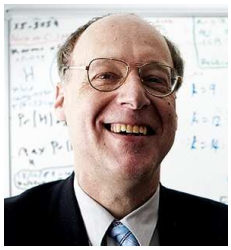
Wilson, M., Hanlon, R. T., Tyack, P. L., and Madsen, P. T. (2007). Intense ultrasonic clicks from echolocating toothed whales do not elicit anti–predator responses or debilitate the squid Loligo pealeii. *Biology letters*, 3(3):225–227.

Yong, E. (November 2012). A large-scale model of the functioning brain. *Nature*, 29.

Zadeh, L. A. (1976). A fuzzy-algorithmic approach to the definition of complex or imprecise concepts. *International Journal of Man-machine studies*, 8(3):249–291.

Zuberbühler, K. (2000). Interspecies semantic communication in two forest primates. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267(1444):713–718.

About the authors



David L. Dowe is currently an Associate Professor in the Clayton School of Information Technology (Computer Science) at Monash University in Melbourne, Australia. He holds a B.Sc. (Hons.) in Mathematics (minor in Physics) from the University of Melbourne and a PhD from Monash. He worked closely with Chris Wallace (1933-2004), the originator (in 1968) of Minimum Message Length (MML), on MML for 13 years, contributing significantly both to the range of applications of MML and to the development of new approximations for message lengths and MML estimators. His innovations include MML Bayesian nets with both continuous and discrete attributes, MML Bayesian nets with latent factors, a proof of the uniqueness of log-loss and Kullback-Leibler divergence in (scoring) evaluation, and redundant Turing machines — and a conjecture about statistical invariance and statistical consistency sometimes being almost exclusive properties of MML. His interests include a broad range of applications from MML, algorithmic information theory and Solomonoff prediction— including human and non-human intelligence, the technological singularity and its repercussions.

José Hernández-Orallo is currently an Associate Professor in the Department of Information Systems and Computation at Technical University of Valencia (UPV, Spain). He holds a B.Sc. and a M.Sc. in Computer Science from UPV, partly completed at the École Nationale Supérieure de l'Électronique et de ses Applications (France), and a Ph.D. in Logic with a doctoral extraordinary prize from the University of Valencia. His academic and research activities have spanned several areas of artificial intelligence, machine learning, data mining and information systems. He has published four books and about a hundred journal articles and conference papers on these topics.