Turing Machines and Recursive Turing Tests

José Hernández-Orallo[1] and **Javier Insa-Cabrera**[2] and **David L. Dowe**[3] and **Bill Hibbard**[4]

Abstract. The Turing Test, in its standard interpretation, has been dismissed by many as a practical intelligence test. In fact, it is questionable that the imitation *game* was meant by Turing himself to be used as a *test* for evaluating machines and measuring the progress of artificial intelligence. In the past fifteen years or so, an alternative approach to measuring machine intelligence has been consolidating. The key concept for this alternative approach is not the Turing *Test*, but the Turing *machine*, and some theories derived upon it, such as Solomonoff's theory of prediction, the MML principle, Kolmogorov complexity and algorithmic information theory. This presents an antagonistic view to the Turing test, where intelligence tests are based on formal principles, are not anthropocentric, are meaningful computationally and the abilities (or factors) which are evaluated can be recognised and quantified. Recently, however, this computational view has been touching upon issues which are somewhat related to the Turing Test, namely that we may need other intelligent agents in the tests. Motivated by these issues (and others), this paper links these two antagonistic views by bringing some of the ideas around the Turing *Test* to the realm of Turing *machines*.

Keywords: Turing Test, Turing machines, intelligence, learning, imitation games, Solomonoff-Kolmogorov complexity.

1 INTRODUCTION

Humans have been evaluated by other humans in all periods of history. It was only in the 20th century, however, that psychometrics was established as a scientific discipline. *Other* animals have also been evaluated by humans, but certainly not in the context of psychometric tests. Instead, comparative cognition is nowadays an important area of research where non-human animals are evaluated and compared. *Machines* —yet again differently— have also been evaluated by humans. However, no scientific discipline has been established for this.

The Turing Test [31] is still the most popular test for machine intelligence, at least for philosophical and scientific discussions. The Turing Test, as a measurement *instrument* and not as a philosophical argument, is very different to the instruments other disciplines use to measure intelligence in a scientific way. The Turing Test resembles a much more customary (and non-scientific) assessment, which happens when humans interview or evaluate other humans (for whatever reason, including, e.g., personnel selection, sports[1] or other competitions). The most relevant (and controversial) feature of the Turing Test is that it takes *humans* as a touchstone to which machines should be compared. In fact, the comparison is not performed by an objective criterion, but assessed by *human* judges, which is not without controversy. Another remarkable feature (and perhaps less controversial) is that the Turing Test is set on an intentionally restrictive interaction channel: a teletype conversation. Finally, there are some features about the Turing Test which make it more general than other kinds of intelligence tests. For instance, it is becoming increasingly better known that programs can do well at human IQ tests [32][8], because ordinary IQ tests only evaluate narrow abilities and assume that narrow abilities accurately reflect human abilities across a broad set of tasks, which may not hold for non-human populations. The Turing test (and some formal intelligence measures we will review in the following section) can test broad sets of tasks.

We must say that Turing cannot be blamed for all the controversy. The purpose of Turing's imitation game [37] was to show that intelligence could be assessed and recognised in a behavioural way, without the need for directly measuring or recognising some other physical or mental issues such as thinking, consciousness, etc. In Turing's view, intelligence can be just seen as a cognitive ability (or property) that some machines might have and others might not. In fact, the standard scientific view should converge to defining intelligence as an ability that some systems: humans, non-human animals, machines —and collectives thereof—, might or might not have, or, more precisely, might have to a larger or lesser degree. This view has clearly been spread by the popularity of psychometrics and IQ tests.[2]

While there have been many variants and extensions of the Turing Test (see [33] or [31] for an account of these), none of them (and none of the approaches in psychometrics and animal cognition, either) have provided a formal, mathematical definition of what in-

[1] DSIC, Universitat Politècnica de València, Spain. email: jorallo@dsic.upv.es

[2] DSIC, Universitat Politècnica de València, Spain. email: jinsa@dsic.upv.es

[3] Clayton School of Information Technology, Monash University, Australia. email: david.dowe@monash.edu

[4] Space Science and Engineering Center, University of Wisconsin - Madison, USA. email: test@ssec.wisc.edu

[1] In many sports, to see how good a player is, we want competent judges but also appropriate team-mates and opponents. Good tournaments and competitions are largely designed so as to return (near) maximal expected information.

[2] In fact, the notion of consciousness and other phenomena is today better separated from intelligence than it was sixty years ago. They are now seen as related but different things. For instance, nobody doubts that a team of people can score well in a single IQ test (working together). In fact, the team, using a teletype communication as in the Turing Test, can dialogue, write poetry, make jokes, do complex mathematics and all these human things. They can even do these things continuously for days or weeks, while some of the particular individuals rest, eat, go to sleep, die, etc. Despite all of this happening *on the other side of the teletype communication*, the system is just regarded as one subject. So the fact that we can effectively measure the cognitive abilities of the team or even make the team pass the Turing Test does not lead us directly to statements such as 'the team has a mind' or 'the team is conscious'. At most, we say this in a figurative sense, as we use it for the *collective consciousness* of a company or country. In the end, the 'team of people' is one of the best arguments against Searle's Chinese room and a good reference whenever we are thinking about evaluating intelligence.

telligence is and how it can be measured.

A different approach is based on one of the things that the Turing Test is usually criticised for: *learning*[3]. This alternative approach requires a proper definition of learning, and actual mechanisms for measuring learning ability. Interestingly, the answer to this is given by notions devised from Turing machines. In the 1960s, Ray Solomonoff 'solved' the problem of induction (and the related problems of prediction and learning) [36] by the use of Turing machines. This, jointly with the theory of inductive inference given by the Minimum Message Length (MML) principle [39, 40, 38, 5], algorithmic information theory [1], Kolmogorov complexity [25, 36] and compression theory, paved the way in the 1990s for a new approach for defining and measuring intelligence based on algorithmic information theory. This approach will be summarised in the next section.

While initially there was some connection to the Turing Test, this line of research has been evolving and consolidating in the past fifteen years (or more), cutting all the links to the Turing Test. This has provided important insights into what intelligence is and how it can be measured, and has given clues to the (re-)understanding of other areas where intelligence is defined and measured, such as psychometrics and animal cognition.

An important milestone of this journey has been the recent realisation in this context that (social) intelligence is the ability to perform well in an environment full of other agents of similar intelligence. This is a consequence of some experiments which show that when performance is measured in environments where no other agents coexist, some important traits of intelligence are not fully recognised. A solution for this has been formalised as the so-called Darwin-Wallace distribution of environments (or tasks) [18]. The outcome of all this is that it is increasingly an issue whether intelligence might be needed to measure intelligence. But this is not because we might need intelligent judges as in the Turing Test, but because we may need other intelligent agents to become part of the exercises or tasks an intelligence test should contain (as per footnote 1).

This seems to take us back to the Turing Test, a point some of us deliberately abandoned more than fifteen years ago. Re-visiting the Turing Test now is necessarily very different, because of the technical companions, knowledge and results we have gathered during this journey (universal Turing machines, compression, universal distributions, Solomonoff-Kolmogorov complexity, MML, re-inforcement learning, etc.).

The paper is organised as follows. Section 2 introduces a short account of the past fifteen years concerning definitions and tests of machine intelligence based on (algorithmic) information theory. It also discusses some of the most recent outcomes and positions in this line, which have led to the notion of Darwin-Wallace distribution and the need for including other intelligent agents in the tests, suggesting an inductive (or recursive, or iterative) test construction and definition. This is linked to the notion of recursive Turing Test (see [32, sec. 5.1] for a first discussion on this). Section 3 analyses the base case by proposing several schemata for evaluating systems that are able to imitate Turing machines. Section 4 defines different ways of doing the recursive step, inspired by the Darwin-Wallace distribution and ideas for making this feasible. Section 5 briefly explores how all this might develop, and touches upon concepts such as universality in Turing machines and potential intelligence, as well as some suggestions as to how machine intelligence measurement might develop in the future.

2 MACHINE INTELLIGENCE MEASUREMENT USING TURING MACHINES

There are, of course, many proposals for intelligence definitions and tests *for machines* which are not based on the Turing Test. Some of them are related to psychometrics, some others may be related to other areas of cognitive science (including animal cognition) and some others originate from artificial intelligence (e.g., some competitions running on specific tasks such as planning, robotics, games, reinforcement learning, . . .). For an account of some of these, the reader can find a good survey in [26]. In this section, we will focus on approaches which use Turing machines (and hence computation) as a basic component for the definition of intelligence and the derivation of tests for machine intelligence.

Most of the views of intelligence in computer science are sustained over a notion of intelligence as a special kind of information processing. The nature of information, its actual content and the way in which patterns and structure can appear in it can only be explained in terms of algorithmic information theory. The Minimum Message Length (MML) principle [39, 40] and Solomonoff-Kolmogorov complexity [36, 25] capture the intuitive notion that there is structure –or redundancy– in data if and only if it is compressible, with the relationship between MML and (two-part) Kolmogorov complexity articulated in [40][38, chap. 2][5, sec. 6]. While Kolmogorov [25] and Chaitin [1] were more concerned with the notions of randomness and the implications of all this in mathematics and computer science, Solomonoff [36] and Wallace [39] developed the theory with the aim of explaining how learning, prediction and inductive inference work. In fact, Solomonoff is said to have 'solved' the problem of induction [36] by the use of Turing machines. He was also the first to introduce the notions of universal distribution (as the distribution of strings given by a UTM from random input) and the invariance theorem (which states that the Kolmogorov complexity of a string calculated with two different reference machines only differs by a constant which is independent of the string).

Chaitin briefly made mention in 1982 of the potential relationship between algorithmic information theory and measuring intelligence [2], but actual proposals in this line did not start until the late 1990s. The first proposal was precisely introduced over a Turing Test and as a response to Searle's Chinese room [35], where the subject was *forced* to learn. This *induction-enhanced* Turing Test [7][6] could then evaluate a general inductive ability. The importance was not that any kind of ability could be included in the Turing Test, but that this ability could be formalised in terms of MML and related ideas, such as (two-part) compression.

Independently and near-simultaneously, a new intelligence test (C-test) [19] [12] was derived as sequence prediction problems which were generated by a universal distribution [36]. The difficulty of the exercises was mathematically derived from a variant of Kolmogorov complexity, and only exercises with a certain degree of difficulty were included and weighted accordingly. These exercises were very similar to those found in some IQ tests, but here they were created from computational principles. This work 'solved' the traditional subjectivity objection of the items in IQ tests, i.e., since the continuation of each sequence was derived from its shortest explanation. However, this test only measured one cognitive ability and its presentation was too narrow to be a general test. Consequently,

[3] This can be taken as further evidence for Turing not conceiving the imitation test as an actual test for intelligence, because the issue about machines being able to learn was seen as inherent to intelligence for Turing [37, section 7], and yet the Turing Test is not especially good at detecting learning ability *during* the test.

these ideas were extended to other cognitive abilities in [14] by the introduction of other 'factors', and the suggestion of using interactive tasks where "rewards and penalties could be used instead", as in reinforcement learning[13].

Similar ideas followed relating compression and intelligence. Compression tests were proposed as a test for artificial intelligence [30], arguing that "optimal text compression is a harder problem than artificial intelligence as defined by Turing's". Nonetheless, the fact that there is a connection between compression and intelligence does not mean that intelligence can be just defined as compression ability (see, e.g., [9] for a full discussion on this).

Later, citeLeggHutter2007 would propose a notion which they referred to as a "universal intelligence measure" —universal because of its proposed use of a universal distribution for the weighting over environments. The innovation was mainly their use of a reinforcement learning setting, which implicitly accounted for the abilities not only of learning and prediction, but also of planning. An interesting point for making this proposal popular was its conceptual simplicity: intelligence was just seen as average performance in a range of environments, where the environments were just selected by a universal distribution.

While innovative, the universal intelligence *measure* [27] showed several shortcomings stopping it from being a viable *test*. Some of the problems are that it requires a summation over infinitely many environments, it requires a summation over infinite time within each environment, Kolmogorov complexity is typically not computable, disproportionate weight is put on simple environments (e.g., with $1-2^{-7} > 99\%$ of weight put on environments of size less than 8, as also pointed out by [21]), it is (static and) not adaptive, it does not account for time or agent speed, etc

Hernandez-Orallo and Dowe [17] re-visited this to give an intelligence *test* that does not have these abovementioned shortcomings. This was presented as an anytime universal intelligence test. The term *universal* here was used to designate that the test could be applied to any kind of subject: machine, human, non-human animal or a community of these. The term *anytime* was used to indicate that the test could evaluate any agent speed, it would adapt to the intelligence of the examinee, and that it could be interrupted at any time to give an intelligence score estimate. The longer the test runs, the more reliable the estimate (the average reward [16]).

Preliminary tests have since been done [23, 24, 28] for comparing human agents with non-human AI agents. These tests seem to succeed in bringing theory to practice quite seamlessly and are useful to compare the abilities of systems of the same kind. However, there are some problems when comparing systems of different kind, such as human and AI algorithms, because the huge difference of both (with current state-of-the-art technology) is not clearly appreciated. One explanation for this is that (human) intelligence is the result of the adaptation to environments where the probability of other agents (of lower or similar intelligence) being around is very high. However, the probability of having another agent of even a small degree of intelligence just by the use of a universal distribution is discouragingly remote. Even in environments where are other agents are included on purpose [15], it is not clear that these agents properly represent a rich 'social' environment. In [18], the so-called Darwin-Wallace distribution is introduced where environments are generated using a universal distribution for multi-agent environments, where a number of agents populate the environment also generated by a universal distribution. The probability of having interesting environments and agents is very low on this first 'generation'. However, if an intelligence test is administered to this population and only those with a certain level are preserved, we may get a second population whose agents will have a slightly higher degree of intelligence. Iterating this process we have different levels for the Darwin-Wallace distribution, where evolution is solely driven (boosted) by a fitness function which is just measured by intelligence tests.

3 THE BASE CASE: THE TURING TEST FOR TURING MACHINES

A recursive approach can raise the odds for environments and tasks of having a behaviour which is attributed to more intelligent agents. This idea of recursive populations can be linked to the notion of *recursive Turing Test* [32, sec. 5.1], where the agents which have succeeded at lower levels could be used to be compared at higher levels. However, there are many interpretations of this informal notion of a recursive Turing Test. The fundamental idea is to eliminate the human reference from the test using recursion —either as the subject that has to be imitated or the judge which is used to tell between the subjects.

Before giving some (more precise) interpretations of a recursive version of the Turing Test, we need to start with the *base case*, as follows (we use TM and UTM for Turing Machine and Universal Turing Machine respectively):

Definition 1 *The imitation game for Turing machines*[4] *is defined as a tuple* $\langle D, B, C, I\rangle$

- *The reference subject A is randomly taken as a TM using a distribution D.*
- *Subject B (the evaluee) tries to emulate A.*
- *The similarity between A and B is 'judged' by a criterion or judge C through some kind of* interaction *protocol I. The test returns this similarity.*

An instance of the previous schema requires us to determine the distribution D and the similarity criterion C and, most especially, how the interaction I goes. In the classical Turing Test, we know that D is the human population, C is given by a human judge, and the interaction is an open teletype conversation[5]. Of course, other distributions for D could lead to other tests, such as, e.g., a canine test, taking D as a dog population, and judges as other dogs which have to tell which is the member of the species or perhaps even how intelligent it is (for whatever purpose —e.g., mating or idle curiosity).

More interestingly, one possible instance for Turing machines could go as follows. We can just take D as a universal distribution over a reference UTM U, so $p(A) = 2^{-K_U(A)}$, where $K_U(A)$ is the prefix-free Kolmogorov complexity of A relative to U. This means that simple reference subjects have higher probability than complex subjects. Interaction can go as follows. The 'interview' consists of questions as random finite binary strings using a universal distribution $s_1, s_2, ...$ over another reference UTM, V. The test starts by subjects A and B receiving string s_1 and giving two sequences a_1 and b_1

[4] The use of Turing machines for the reference subject is relevant and not just a way to link two things by their name, Turing. Turing machines are required because we need to define formal distributions on them, and this cannot be done (at least theoretically) for humans, or animals or 'agents'.

[5] This free teletype conversation may be problematic in many ways. Typically, the judge C wishes to steer the conversation in directions which will enable her to get (near-)maximal (expected) information (before the time-limit deadline of the test) about whether or not the evaluee subject B is or is not from D. One tactic for a subject which is not from D (and not a good imitator either) is to distract the judge C and steer the conversation in directions which will give judge C (near-) minimal (expected) information.

as respective answers. Agent B will also receive what A has output immediately after this. Judge C is just a very simple function which compares whether a_1 and b_1 are equal. After one interation, the system issues string s_2. After several iterations, the score (similarity) given to B is calculated as an aggregation of the times a_i and b_i have been equal.

This can be seen as formalisation of the Turing Test where it is a Turing machine that needs to be imitated, and the criterion for imitation is the similarity between the answers given by A and B to the same questions. If subject B cannot be told or instructed about the goal of the test (imitating A) then we can use rewards after each step, possibly concealing A's outputs from B as well.

This test might seem ridiculous at first sight. Some might argue that being able to imitate a randomly-chosen TM is not related to intelligence. However, two issues are important here. First, agent B does not know who A is in advance. Second, agent B tries to imitate A solely from its behaviour.

This makes the previous version of the test very similar to the most abstract setting used for analysing what learning is, how much complexity it has and whether it can be solved. First, this is tantamount to Gold's language identification in the limit [11]. If subject B is able to identify A at some point, then it will start to score perfectly from that moment. While Gold was interested in whether this could be done in general and for every possible A, here we are interested in how well B does this on average for a randomly-chosen A from a distribution. In fact, many simple TMs can be identified quite easily, such as those simple TMs which output the same string independently of the input. Second, and following this averaging approach, Solomonoff's setting is also very similar to this. Solomonoff proved that B could get the best estimations for A if B used a mixture of all consistent models inversely weighted by 2 to the power of their Kolmogorov complexity. While this may give the best theoretical approach for prediction and perhaps for "imitation", it does not properly "identify" A. Identification can only be properly claimed if we have one single model of A which is exactly as A. This distinction between one vs. multiple models is explicit in the MML principle, which usually considers just one single model, the one with the shortest two-part message encoding of said model followed by the data given this model.

There is already an intelligence test which corresponds to the previous instance of definition 1, the C-test, mentioned above. The C-test measures how well an agent B is able to identify the pattern behind a series of sequences (each sequence is generated by a different program, i.e., a different Turing machine). The C-test does not use a query-answer setting, but the principles are the same.

We can develop a slight modification of definition 1 by considering that subject A also tries to imitate B. This might lead to easy convergence in many cases (for relatively intelligent A and B) and would not be very useful for comparing A and B effectively. A significant step forward is when we consider that the goal of A is to make outputs that cannot be imitated by B. While it is clearly different, this is related to some versions of Turing's imitation game, where one of the human subjects pretends to be a machine. While there might be some variants here to explore, if we restrict the size of the strings used for questions and answers to 1 (this makes agreeing and disagreeing equally likely), this is tantamount to the game known as 'matching pennies' (a binary version of rock-paper-scissors where the first player has to match the head or tail of the second player, and the second player has to disagree on the head or tail of the first). Interestingly, this game has also been proposed as an intelligence test in the form of Adversarial Sequence Prediction [20][22] and is related to the "elusive model paradox" [3, footnote 211][4, p 455][5, sec. 7.5].

This instance makes it more explicit that the distribution D over the agents that the evaluee has to imitate or compete with is crucial. In the case of imitation, however, there might be non-intelligent Turing machines which are more difficult to imitate/identify than many intelligent Turing machines, and this difficulty seems to be related to the Kolmogorov complexity of the Turing machine. And linking difficulty to Kolmogorov complexity is what the C-test does. But biological intelligence is frequently biased to social environments, or at least to environments where other agents can be around eventually. In fact, societies are usually built on common sense and common understanding, but in humans this might be an evolutionarily-acquired ability to imitate other humans, but not other intelligent beings in general. Some neurobiological structures, such as *mirror neurons* have been found in primates and other species, which may be responsible of understanding what other people do and will do, and for learning new skills by imitation. Nonetheless, we must say that human unpredictability is frequently impressive, and its relation to intelligence is far from being understood. Interestingly, some of the first analyses on this issue [34][29] linked the problem with the competitive/adversarial scenario, which is equivalent to the matching pennies problem, where the intelligence of the peer is the most relevant feature (if not the only one) for assessing the difficulty of the game, as happens in most games. In fact, matching pennies is the purest and simplest game, since it reduces the complexity of the 'environment' (rules of the game) to a minimum.

4 RECURSIVE TURING TESTS FOR TURING MACHINES

The previous section has shown that introducing agents (in this case, agent A) in a test setting requires a clear assessment of the distribution which is used for introducing them. A general expression of how to make a Turing Test for Turing machines recursive is as follows:

Definition 2 *The recursive imitation game for Turing machines is defined as a tuple $\langle D, C, I\rangle$ where tests and distributions are obtained as follows:*

```
1. Set $D_0 = D$ and $i = 0$.
2. For each agent B in a sufficiently large set of TMs
3.     Apply a sufficiently large set of instances of definition 1 with parameters $\langle D_i, B, C, I\rangle$.
4.     B's intelligence at degree $i$ is averaged from this sample of imitation tests.
5. End for
6. Set $i = i + 1$
7. Calculate a new distribution $D_i$ where each TM has a probability which is directly related to its intelligence at level $i - 1$.
8. Go to 2
```

This gives a sequence of D_i.

The previous approach is clearly uncomputable in general, and still intractable even if reasonable samples, heuristics and step limitations are used. A better approach to the problem would be some kind of propagation system, such as Elo's rating system of chess [10], which has already been suggested in some works and competitions in artificial intelligence. A combination of a *soft* universal distribution, where simple agents would have slightly higher probability, and a one-vs-one credit propagation system such as Elo's rating (or any other mechanism which returns maximal expected information with a minimum of pairings), could feasibly aim at having a reasonably good estimate of the relative abilities of a big population of Turing machines, including some AI algorithms amongst them.

What would this rating mean? If we are using the imitation game, a high rating would show that the agent is able to imitate/identify other agents of lower rating well and that it is a worse imitator/identifier than other agents with higher rating. However, there is no reason to think that the relations are transitive and anti-reflexive; e.g., it might even happen that an agent with very low ranking would be able to imitate an agent with very high ranking better than the other way round.

One apparently good thing about this recursion and rating system is that the start-up distribution can be very important from the point of view of heuristics, but it might be less important for the final result. This is yet another way of escaping from the problems of using a universal distribution for environments or agents, because very simple things take almost all the probability —as per section 2. Using difficulty as in the C-test, making adaptive tests such as the anytime test, setting a minimum complexity value [21] or using hierarchies of environments [22] where "an agent's intelligence is measured as the ordinal of the most difficult set of environments it can pass" are solutions for this. We have just seen another possible solution where evaluees (or similar individuals) can take part in the tests.

5 DISCUSSION

The Turing test, in some of its formulations, is a game where an agent tries to imitate another (or its species or population) which might (or might not) be cheating. If both agents are fair, and we do not consider any previous information about the agents (or their species or populations), then we have an imitation test for Turing machines. If one is cheating, we get closer to the adversarial case we have also seen.

Instead of including agents arbitrarily or assuming that any agent has a level of intelligence a priori, a recursive approach is necessary. This is conceptually possible, as we have seen, although its feasible implementation needs to be carefully considered, possibly in terms of rankings after random 1-vs-1 comparisons.

This view of the (recursive) Turing test in terms of Turing machines has allowed us to connect the Turing test with fundamental issues in computer science and artificial intelligence, such as the problem of learning (as identification), Solomonoff's theory of prediction, the MML principle, game theory, etc. These connections go beyond to other disciplines such as (neuro-)biology, where the role of imitation and adversarial prediction are fundamental, such as predator-prey games, mirror neurons, common coding theory, etc. In addition, this has shown that the line of research with intelligence tests derived from algorithmic information theory and the recent Darwin-Wallace distribution are also closely related to this as well. This (again) links this line of research to the Turing test, where humans have been replaced by Turing machines.

This sets up many avenues for research and discussion. For instance, the idea that the ability of imitating relates to intelligence can be understood in terms of the universality of a Turing machine, i.e. the ability of a Turing machine to emulate another. If a machine can emulate another, it can acquire all the properties of the latter, including intelligence. However, in this paper we have referred to the notion of 'imitation', which is different to the concept of Universal Turing machine, since a UTM is defined as a machine such that there is an input that turns it into any other pre-specified Turing machine. A machine which is able to imitate well is a good learner, which can finally identify any pattern on the input and use it to imitate the source. In fact, a good imitator is, *potentially*, very intelligent, since it can, in theory (and disregarding efficiency issues), act as any other very intelligent being by just observing its behaviour. Turing advocated for learning machines in section 7 of the very same paper [37] where he introduced the Turing Test. Solomonoff taught us what learning machines should look like. We are still struggling to make them work in practice and preparing for assessing them.

ACKNOWLEDGEMENTS

This work was supported by the MEC projects EXPLORA-INGENIO TIN 2009-06078-E, CONSOLIDER-INGENIO 26706 and TIN 2010-21062-C02-02, and GVA project PROMETEO/2008/051. Javier Insa-Cabrera was sponsored by Spanish MEC-FPU grant AP2010-4389.

REFERENCES

[1] G. J. Chaitin, 'On the length of programs for computing finite sequences', *Journal of the Association for Computing Machinery*, **13**, 547–569, (1966).

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

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

[4] D. L. Dowe, 'Minimum Message Length and statistically consistent invariant (objective?) Bayesian probabilistic inference - from (medical) "evidence"', *Social Epistemology*, **22**(4), 433 – 460, (October - December 2008).

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

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

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

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

[9] D. L. Dowe, J. Hernández-Orallo, and P. K. Das, 'Compression and intelligence: social environments and communication', in *Artificial General Intelligence*, eds., J. Schmidhuber, K.R. Thórisson, and M. Looks, volume 6830, pp. 204–211. LNAI series, Springer, (2011).

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

[11] E.M. Gold, 'Language identification in the limit', *Information and control*, **10**(5), 447–474, (1967).

[12] J. Hernández-Orallo, 'Beyond the Turing Test', *J. Logic, Language & Information*, **9**(4), 447–466, (2000).

[13] J. Hernández-Orallo, 'Constructive reinforcement learning', *International Journal of Intelligent Systems*, **15**(3), 241–264, (2000).

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

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

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

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

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

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

[20] B. Hibbard, 'Adversarial sequence prediction', in *Proceeding of the 2008 conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, pp. 399–403. IOS Press, (2008).

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

[22] B. Hibbard, 'Measuring agent intelligence via hierarchies of environments', *Artificial General Intelligence*, 303–308, (2011).

[23] J. Insa-Cabrera, D. L. Dowe, S. España-Cubillo, M. Victoria Hernández-Lloreda, and José Hernández-Orallo, 'Comparing humans and ai agents', in *AGI: 4th Conference on Artificial General Intelligence - Lecture Notes in Artificial Intelligence (LNAI)*, volume 6830, pp. 122–132. Springer, (2011).

[24] J. Insa-Cabrera, D. L. Dowe, and José Hernández-Orallo, 'Evaluating a reinforcement learning algorithm with a general intelligence test', in *CAEPIA - Lecture Notes in Artificial Intelligence (LNAI)*, volume 7023, pp. 1–11. Springer, (2011).

[25] A. N. Kolmogorov, 'Three approaches to the quantitative definition of information', *Problems of Information Transmission*, **1**, 4–7, (1965).

[26] S. Legg and M. Hutter, 'Tests of machine intelligence', in *50 years of artificial intelligence*, pp. 232–242. Springer-Verlag, (2007).

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

[28] S. Legg and J. Veness, 'An Approximation of the Universal Intelligence Measure', in *Proceedings of Solomonoff 85th memorial conference*. Springer, (2012).

[29] D. K. Lewis and J. Shelby-Richardson, 'Scriven on human unpredictability', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, **17**(5), 69 – 74, (October 1966).

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

[31] G. Oppy and D. L. Dowe, 'The Turing Test', in *Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta. Stanford University, (2011). http://plato.stanford.edu/entries/turing-test/.

[32] P. Sanghi and D. L. Dowe, 'A computer program capable of passing IQ tests', in *4th Intl. Conf. on Cognitive Science (ICCS'03), Sydney*, pp. 570–575, (2003).

[33] A.P. Saygin, I. Cicekli, and V. Akman, 'Turing test: 50 years later', *Minds and Machines*, **10**(4), 463–518, (2000).

[34] M. Scriven, 'An essential unpredictability in human behavior', in *Scientific Psychology: Principles and Approaches*, eds., B. B. Wolman and E. Nagel, 411–425, Basic Books (Perseus Books), (1965).

[35] J. R. Searle, 'Minds, brains and programs', *Behavioural and Brain Sciences*, **3**, 417–457, (1980).

[36] R. J. Solomonoff, 'A formal theory of inductive inference', *Information and Control*, **7**, 1–22, 224–254, (1964).

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

[38] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*, Information Science and Statistics, Springer Verlag, May 2005. ISBN 0-387-23795X.

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

[40] C. S. Wallace and D. L. Dowe, 'Minimum message length and Kolmogorov complexity', *Computer Journal*, **42**(4), 270–283, (1999).