Towards a Staged Developmental Intelligence Test for Machines

Ed Keedwell
College of Engineering, Mathematics and Physical Sciences, University of Exeter, Harrison Building, North Park Road, Exeter, EX4 4QF {E.C.Keedwell@ex.ac.uk}

Abstract. The Turing Test provides a test for determining machine intelligence that has proven to be very difficult to overcome for researchers in AI, perhaps because of its pass/fail nature. A new test is proposed here, known as the Staged Developmental Machine Intelligence test which uses a number of stages based on the testing of development in children as its basis. It is proposed that this test would reduce the need for sophisticated natural language processing and provide a framework for evaluating machines on a scale rather than providing a binary result as with the original Turing Test.

1 INTRODUCTION

Some 50 years ago, Alan Turing postulated a test of machine intelligence, the Imitation Game [7] that has been generally been accepted as the benchmark that machines will be required to achieve before they can be considered intelligent. The test, now widely known as the Turing Test is in essence very simple and has been extensively described in the literature so a short description will suffice here. A human arbiter uses some form of communication (often this is described as a computer terminal) to speak to two individuals that are located elsewhere. The original task of each individual in the Imitation Game is to convince the arbiter that they are female, although further interpretations of the test consider that each individual attempts to convince the arbiter that the other is a machine. If the machine successfully convinces the arbiter that it is human, then it has passed the Turing Test and can be said to be intelligent. An important factor in this test is that the human is never in direct contact with the other individuals and so cannot make inferences based on appearance or other subconscious prejudices.

Replies to the Turing Test

As might be expected with a test of such high profile, there have been numerous replies and criticisms of this test. One of the most famous is that of the Chinese Room Argument posed by John Searle [8] in 1980. Searle states that the appearance of intelligent behaviour as required by the Turing Test is not sufficient for the machine to be considered intelligent as it is simply manipulating symbols and is not able to attach meaning to these symbols – the so called *symbol grounding problem*. A number of responses followed and Searle replied to each of these responses with largely the same notion that the symbols, whether learnt or pre-programmed, representing language, images or neural network weights, are all essentially still symbols and suffer from the same symbol grounding problem. The test proposed here is less reliant on the processing of symbolic information and so may be less prone to the arguments of Searle than the original test, although elements of the Robot Reply will likely still hold.

Issues with the Turing Test

It is proposed here that the main issue with the Turing Test is that it has a binary outcome. A system can either pass the test and be considered intelligent or fail it, with no possibility for a system to pass a portion of the test. Recent implementations of the test such as the Loebner Prize[1] have necessarily awarded a degree of 'intelligence' to the submissions as no machines has yet passed the Turing Test. The approach is to use a number of judges and a ranking system, where the winner of the prize is the system which convinces the most judges that they are most human. In short, the binary nature of the Turing Test might seem to be advantageous when considering if a system is intelligent, but as shown by the modifications utilised by the Loebner prize, this idea does not fit very well with the way that humans this about intelligence.

Degrees of Intelligence

In language, we humans often describe animals, adult humans and children as possessing a certain level of intelligence, effectively expressing intelligence as a scalar value. In particular, we discuss the level of intelligence in at least three different ways:

1. **The level of intelligence possessed by lower organisms.** This point can be characterised as a series of questions: Is a plant intelligent? Not by many standards, but it displays many more characteristics of intelligence (e.g. the ability to align itself to stimuli) than inanimate objects, a point Turing makes in his original paper. What about bacteria? Many of these have flagella or other methods of propulsion and so can seek out food or move away from danger. What about cats and dogs? Well perhaps obviously they are closer to the sort of intelligence we mean when testing machines for 'intelligence' as they can recognise patterns, have memory, can be conditioned to stimuli etc... Behaviour that we recognise as some of the more basic functions of our own intellect.
2. **The level of intelligence possessed by adult humans.** A key actor in the Turing Test is the human arbiter who makes the decision as to whether the machine is intelligent. However, surely it would be easier to convince some individuals than others? Of course we often discuss the level of intelligence or otherwise of individuals within a community, and have developed our own scalar methods of

[1] http://www.loebner.net/Prizef/loebner-prize.html

assessing intelligence through IQ tests. These tests have somewhat of a chequered past which is not within the scope of this argument, but it is clear that intelligence is considered to be a scalar property of a human, not a binary measure.

3. **The level of intelligence possessed by developing children.** The development of intelligent behaviour takes some considerable time in human children, especially the development of language. Piaget [5] postulated a four stage process of learning in his constructivist framework for cognitive development. Others have since expressed this in terms of a continuum, and whichever of these models holds, it is clear that there are some cognitive functions that have to be learned before others can then be learned. Implicit in these theories is the notion that a child does not suddenly become intelligent, but that there is a gradual building up of the faculties necessary for intelligence, again hinting at a scalar property.

These three familiar examples hopefully convince that we as humans do not consider intelligence to be a binary property, even among the adults of our own species. So if we accept that there is variation in the intelligence of our own adult population, why do we persist with a binary test for machines? A more empirical test should surely relate to the ways in which we judge ourselves and other organisms on their intelligence. In short, a Staged Developmental Intelligence Test. Aside from being aligned more closely to our own descriptions of organisms possessing levels of intelligence, a Staged Developmental Machine Intelligence Test would have the added benefit that we would be able to compare the intelligence of machines without resorting to the rather subjective ranking and repeated measures currently used with implementations of the Turing Test.

Developmental Alternatives to the Turing Test

Many alternatives have been proposed to the test since Turing originally proposed his original test in 1950 (e.g. the text compression test [4]). I will not attempt to discuss them all here, and readers are directed to [3] for a review of alternatives to the test. However, there are a handful of these tests which are relevant here. Firstly, in [6] the authors propose a test based on the acquisition of language and relate the complexity of the acquired vocabulary to the intelligence of the machine. This approach establishes the key aspect of a developmental intelligence test, the notion that machines can possess degrees of intelligence. However, the test itself is still highly dependent on the language elements of intelligence and therefore the field of natural language processing. An alternative test, known as the 'Toddler Turing Test' in [1] describes a test similar in nature to that proposed here, but only takes into account one stage of development as proposed by Piaget. The approach described here extends the idea of [1] to further stages of cognitive development without the reliance on text processing of the test shown in [6].

2 COGNITIVE DEVELOPMENT

The ideas of learning and cognitive development are closely associated with the development of mature intelligence that might be tested by the Turing Test. Many theories exist as to how human infants learn about the world but perhaps the most widely known are those of Piaget. Piaget's constructivist view [5] of development contrasted starkly with the opposing behaviourist and nativist approaches that existed at the time. His ideas were among the first to suggest that both nature and nurture play a part in the cognitive development of children. Although more recent theories have superseded Piaget's work somewhat, his ideas remain highly influential in many areas of education including education policy. The principles of testing at 7 and 11 years of age are as result of his staged approach to development. More recent theories (notably Siegler [9][10]) have shed some doubt as to the validity of a staged approach, suggesting that there is overlap between the stages, but the Staged Development Intelligence Test suggested here could be adapted to new development schemes as they are put forward.

Piaget's Stages of Cognitive Development

Piaget's theory [5] states that cognitive development occurs in four stages, during which a child learns new concepts. Importantly, if a child is in one stage, the theory states that they cannot learn concepts from another, more advanced stage. The four stages are shown in Figure 1:

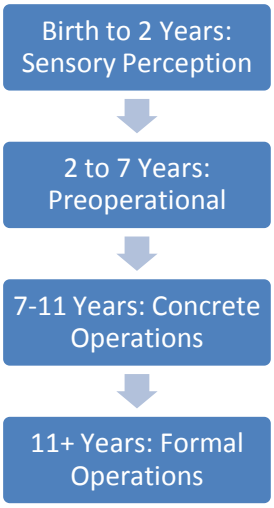

Figure 1 - Piaget's Stages of Cognitive Development

Sensory Perception

In this stage infants are able to distinguish themselves from others and begin to act intentionally on their environment and are able to begin to recognise object permanence. However the actions available are very much dictated by the current stimulus, infants in this stage for instance will smile at a parent's face but cannot intentionally recall that image and smile again.

Pre-Operational

In this stage, children are able to use language to represent objects and to classify objects into classes according to common attributes (e.g. colour). However children in this stage still find impossible to consider other people's points of view and are therefore highly egocentric.

Concrete Operational

Children can think logically about objects and events and can begin to conserve properties of objects (i.e. number (age 6), mass

(age 7) and weight (age 9). Also more advanced classification and sorting behaviours are demonstrated in this stage.

Formal Operational
At this stage children can apply logic to abstract events and objects and can test various hypotheses about the world. The child can also consider hypothetical scenarios, including those that might occur in the future.

It is clear from this framework that full adult intelligent reasoning is built from the ground up. Certain characteristics of the world must be learned before other, more sophisticated concepts can be considered. This is perhaps most clearly demonstrated by the idea of objects – firstly their permanence in the world must be established and then they can be classified according to certain properties they possess. Some of these properties are then learned to be conserved, no matter what transformation is applied to them. Finally objects can be considered in abstract situations and the consequences of actions upon them can be visualised rather than needing to be carried out.

3 IMPLICATIONS FOR THE TURING TEST

This developmental framework above shows that the original Turing Test, due largely to its generality, tests all of the stages of cognitive development (and therefore what we understand as intelligent behaviour), up to and including formal operations. The sorts of questions asked by judges of the Turing Test tend to be experiential in nature, so questions such as "where are you from?" and "have you any plans for later in the day?" are commonplace in the transcripts of competitions such as the Loebner prize.
So whereas we might reasonably expect fully grown adults to be able to answer such questions competently, we wouldn't expect 2 or 3 year olds to engage in conversations that involve these higher concepts of abstract thought. The concepts are quite abstract in nature and therefore test the latter stages of the developmental framework as shown above.

Certainly, as a gold-standard test of whether machine intelligence exists, the Turing Test provides us with a simple-to-implement and robust evaluation of machine intelligence. However, the test itself is not very helpful in the development of a machine intelligence that might one day be able to pass it. By posing such a wide and apparently insurmountable problem to researchers, perhaps we are encouraging the sort of 'tuned' approaches and flawed responses that Turing sought to reduce. The contention here is that we should not be rewarding AI approaches that get marginally closer to the prize by becoming infinitesimally more able to respond in somewhat more human-like terms without materially improving on the intelligence behind the approach.

4 A STAGED DEVELOPMENT TURING TEST

A test of intelligence should be one that measures machine intelligence in broadly the same way as we measure a child's, through a staged approach to cognitive development. Suggestions for the tests are given below. In each case, the human behaviour is first described and is followed by a possible machine test to determine whether the machine has reached this level of capability. The tests are intended to be as generic as possible, but it is perhaps easiest to see how this might work best with a system that processes visual information and can make changes to the environment through robotic (or simulated robotic) action.

Stage 1 – Sensory Perception Stage
In this stage the test would initially involve simple stimulus-response tasks that are conducted with infants. There are a number of these used to assess progression from Birth through to 2 years.

Reacts to basic stimuli – light/sudden sound
- The machine is able to 'attend' to important stimuli (e.g. the facial recognition systems available in some cameras).
- The system will attend to sharp changes in the stimulus.

Understands cause and effect
- The machine can predict the likely movement of objects when the video stimulus is stopped (e.g. a bouncing ball hanging in mid-air). The accuracy of this prediction can be used to determine the degree of proficiency of the machine.

Understands the concept of objects and what to expect from them
- The machine can differentiate between objects of different types (essentially a classification task). In particular, those objects that it can directly effect (e.g. move) and those that it cannot.

Uses trial-and-error to learn about the world
- The machine possesses the ability to modify its internal structure based on experimentation with the world.

Stage 2 – Pre-Operational
The development of language skills comes to the fore in this testing stage. Systems that would pass this stage would be able to demonstrate:

Language understanding relating to itself (but not wider topics)
- The machine should be able to answer questions about itself and the current state it is in. Although still a sophisticated NLP task, the reduction in complexity over the standard Turing Test should make this restricted goal more achievable.

Can relate to objects through language even though they are not present in the perceptual field of the system
- The machine would be required to remember (requiring memory) information about objects that it has 'seen' previously.

Stage 3 – Concrete Operations
Higher order thinking starts to become evident in this phase. Systems that would pass this phase would demonstrate
Conservation of volume, mass etc..
- In this test, the machine should be able to determine the correct quantity of objects despite their arrangement in the visual stimulus.

Classification of objects based on logical basis rather than superficial object characteristics
- The machine should be able to separate objects into logical groups (e.g. animals, shapes etc..)

Sorting of objects

- Given a set of objects, the machine should be able to order them according to some criterion (e.g. size or colour).

Reversibility

- When shown an action taking place (e.g. placing a weight on some scales), the machine should be able to accurately predict what would happen if that action were reversed (e.g. the weight was removed).

Stage 4 – Formal Operations

Logical thinking becomes evident in this phase. Systems that would pass this phase would demonstrate:

The ability to create its own hypotheses

- The machine should no longer be directed by the learning process but should be able to conduct its own experiments to test those hypotheses. (Machines already exist that are able to conduct this hypothesis testing in the restricted field of science [2]).

Abstract Thought

- The machine should be capable of considering stimuli not presented to the machine and should be able to consider the interactions of objects, seen or unseen, in novel ways.

Final Stage

A final stage could be implemented whereby the existing Turing Test (or a visual version of the same using objects rather than language) could be used to ultimately determine that an intelligent machine has been created.

Machines can then be judged according to their effective 'stage' from the framework above. Machines that are capable of all concepts within a stage could then win that particular stage and a new competition would be opened for the next stage. Further agreement and standardisation of the tests would need to be undertaken before it would be ready to be developed as a test in its own right and the test could be adapted for other theories of cognitive development (e.g. Siegler's overlapping waves[9][10]) if required. However, it is the principle of a staged approach is the important contribution here.

A further consideration is that it may be that even the capabilities of the first stages would be too complex for current machines and therefore a finer-grained test could be conducted which adheres to a well known developmental tests (e.g. "Schedule of growing skills", Denver 2, Griffiths and Mary Sheridan developmental tests) which typically test children to a maximum of age 8. These tests all work within the first two stages of Piaget's framework and so would provide a more detailed rating system for machines in these stages.

5 DISCUSSION

This developmental framework shown here has a number of advantages. Firstly, the stage reduces the focus on natural-language processing, a very difficult field of AI which is often the stumbling block for machines wishing to pass the Turing Test and restricts the AI approaches that can be taken to passing the Turing Test to those that can efficiently process symbols. This could lead to more machines (e.g. Asimo) which focus on the perceptual aspects of intelligence rather than the language-based elements, particularly if the visual stimulus tests were conducted as shown in section 4.

Secondly, aspects of intelligence would be investigated in order, leading to a developmentally focussed approach to delivering machine intelligence. For instance object permanence and the ability to manipulate objects would come at an early stage in the process and would be later built upon. This developmental aspect means that we would create machines that more closely mimic our own progress towards intelligence rather than trying to create an intelligent machine from scratch. We tend to associate intelligence with the final product of development, i.e. the adult human, but the process of development may be crucial in the creation of intelligent machines.

Thirdly this test provides an established framework within which to classify machines as to how 'intelligent' they are. Currently machines can only be classified as passing or failing the Turing Test, of which only machines in the latter category exist at present. The impasse that has existed for 60 years would be lifted somewhat as progress would be measurable, even though the ultimate goal remains the same.

However, there will be criticisms of this approach and in the spirit of Turing's original paper some possible replies to the Staged Approach are presented here. Firstly, more than ever, this test would doom us to creating intelligence in our own image. This could certainly be interpreted as a negative aspect as replicating human intelligence processes in a machine might limit the possibilities of what can be achieved with silicon. To counter this I would point to the fact that currently we cannot come anywhere close to replicating our own intelligence, let alone create a new species of intelligent behaviour. A pragmatic way forward would be to create thinking machines in our own image that can pass all tests and only then consider how we might allow the machines to develop differently. Quite simply, if our own intelligence is all we have as a basis for intelligent machines, then shouldn't replicating that be our first priority? In any case, there is no guarantee that creating a human developmental process within a computing machine with its vastly different architecture and capabilities would inevitably lead to human-like intelligence.

Secondly, a related reply might be that the method of achieving intelligence should be independent of any test, perhaps an intelligent system might exist that does not need to develop in the stages shown here. My response to this is that a key point of this test is that it is based on the behaviour of the system in a similar way to the Turing Test and as such is independent of the implementation used. Also it does not prescribe that the system should necessarily develop in this way, but crucially we would expect a system that was capable of passing the final stage would also be capable of passing the previous stages, in the same way that one would expect a 11 year old child to be able to count and conserve volume as well as be capable of abstract, logical thought.

6 CONCLUSION

A Staged Development Intelligence Test has been presented with the main argument that if we are to judge intelligent machines by our own standards of intelligence, then we can enrich the process of AI development by considering degrees of intelligence

displayed during cognitive development. The stages can be considered as milestones towards ultimately solving the original Turing Test, or similar variant, as the final stage in the development of thinking machines.

7 ACKNOWLEDGEMENTS

Thanks to Lynda Keedwell for some informative discussions on child development.

8 REFERENCES

[1] Alvarado, N., Adams, S., Burbeck, S., Latta, C., (2002) "Beyond the Turing Test: Performance Metrics for Evaluating a Computer Simulation of the Human Mind" in *Performance Metrics for Intelligent Systems2002*

[2] King *et al.* (2009) "The Automation of Science" *Science* Vol. **324,** No. 5923, pp 85-89

[3] Legg, S., Hutter, M. (2007) "Tests of Machine Intelligence" in *50 Years of Computer Science LNCS* **4850**, pp232-242

[4] Mahoney, M (1999) "Text Compression as a Test for Artificial Intelligence" in *Proceedings of AAAI/IAAI*

[5] Piaget, J., (1964) "Cognitive Development in Children", *Journal of Research in Science Teaching*, **2**, No 3., pp176-186

[6] Triesten-Goran, A., Dunietz, J., Hutchens, J.L., (2000) "The developmental approach to evaluating artificial intelligence" in *Performance Metrics for Intelligent Systems 2000*

[7] Turing A., (1950) "Computing Machinery and Intelligence" *Mind* **59**, pp433-460

[8] Searle, J., (1980) "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, 417-424.

[9] Siegler, R.S., (2000) "The Rebirth of Children's Learning" *Child Development*, **71**, No 1., pp26-35

[10] Shrager, J., Siegler. R.S., (1998) "SCADS: A Model of Children's Strategy Choices and Strategy Discoveries" *Psychological Science*, **9**, pp405-410