A Non-Behavioural, Computational Extension to the Turing Test

David L. Dowe
Department of Computer Science, Monash University,
Clayton, Vic. 3168, Australia

Alan R. Hájek
HSS, California Institute of Technology, Pasadena,
California 91125, U.S.A.
e-mail: {dld@cs.monash.edu.au, ahajek@hss.caltech.edu}

Abstract

The purely behavioural nature of the Turing Test leaves many with the view that passing it is not sufficient for 'intelligence' or 'understanding'. We propose here an additional necessary computational requirement on intelligence that is non-behavioural in nature and which we contend is necessary for a commonsense notion of 'inductive learning' and, relatedly, of 'intelligence'. Said roughly, our proposal is that a key to these concepts is the notion of compression of data. Where the agent under assessment is able to communicate, e.g. by a tele-type machine, our criterion is that, in addition to requiring the agent's being able to pass Turing's original (behavioural) Turing Test, we also require that the agent have a somewhat compressed representation of the test domain. Our reason for adding this requirement is that, as we shall argue from both Bayesian and information-theoretic grounds, inductive learning and compression are tantamount to the same thing. We can only compress data when we learn a pattern or structure, and it seems quite reasonable to require that an 'intelligent' agent can inductively learn (and record the result learnt from the compression). We illustrate these ideas and our extension of the Turing Test via Searle's Chinese room example.

We also ask the following question: Given two programs H_1 and H_2 respectively of lengths l_1 and l_2, $l_1 < l_2$, if H_1 and H_2 perform equally well (to date) on a Turing Test, which, if either, should be preferred for the future?

We also set a challenge. If humans can presume intelligence in their ability to set the Turing test, then we issue the additional challenge to researchers to get machines to *administer* the Turing Test.

Keywords: Turing Test, Artificial Intelligence, Philosophy of AI, Machine Learning, Bayesian and Statistical Learning Methods, Cognitive Modelling.

1 Introduction - the Turing Test and Chinese Room

Turing introduced his famous test, "the Turing Test"[10], of (artificial) intelligence by proposing that the agent be tested for the ability to simulate by tele-type the conversational actions of a human[a]. One possible way for a machine to carry out such a simulation would be for it to be programmed with a list of possible remarks that the human tester might make and corresponding recommended responses for the machine to generate in each case. The conversation could be thought of as developing along the lines of a game tree, with moves alternating between human and machine: the machine has to generate a satisfactory response at every point in the game tree that the succession of remarks leads it to, and the human tries to catch the machine out (or concedes that (s)he can't catch it out). Searle[8] gives a parallel example in which, instead of a machine trying to simulate humanness, a human endeavours to simulate an operational understanding of Chinese[b]. This involves a human operator with no knowledge of Chinese other than a look-up table, replying to input strings of Chinese characters with chosen output strings of Chinese characters. Among other things, Searle asks us to consider the case of an operator who memorises the look-up table.

Behaviourally, the operator who has memorised the look-up table will pass the Turing test for understanding Chinese. Our objection to the Turing test and our consequent proposed non-behavioural enhancement are based on our belief that understanding a subject domain has something to do with the compression of relevant data. This objection and enhancement are perhaps best highlighted by comparing the computational resources available to an English-speaking[c] human who also speaks Chinese with the computational resources required by an English-speaking human to store and access such a look-up table. Whatever might constitute sufficient conditions for a commonsense notion of "intelligence", we contend that, as well as an ability to pass the Turing Test, it is also *necessary* to have a compression of the relevant test subject matter. The greater the compression, typically, the greater the understanding.

The reader is referred to an expanded version of this paper[3] for additional detail, including touching on the problem of other minds.

[a] This work was supported by Australian Research Council (ARC) Large Grants Nos. A49330656 and A49602504.

[b] we follow Searle in not specifying which particular dialect of Chinese.

[c] clearly, the language need not be English. Any language sufficiently different to "Chinese" would suffice.

2 Inductive Learning = compression

We wish to put forward the view that learning from some body of data is typically an act of compression of that data. Such a theory has been explicitly stated elsewhere[14] for learning languages, but we wish to propose it for all inductive learning. The idea of using notions of compression to carry out statistical and inductive inference was suggested in the 1960s[9,2,11] and has been successfully implemented in Minimum Message Length (MML)[11,13] and Minimum Description Length (MDL)[7] applications ever since, both of which are related to Kolmogorov complexity[6,4]. For the reader possibly unfamiliar[d] with MML and MDL[e], consider firstly a set of data involving two variables, x_1 and x_2 (as it might be, force and acceleration). We begin with a long data string consisting of ordered pairs of the form (x_1, x_2). One way to summarize the data is simply to record this string. However, suppose we notice that apparently $x_2 \approx kx_1$ – that is, the data points all lie on or near a straight line of slope k. Then the data string can be compressed without loss : we can record instead just the x_1 values, this functional relationship, and some error terms. Moreover, we have increased our understanding of the data by (inductively) learning this relationship. To be sure, there are other candidate functions for the relationship between x_1 and x_2. What we want, however, is the function that gives the greatest amount of compression, – the minimum message (or description) length (MML, or MDL) encoding of the data[f].
We believe that these points generalize. Understanding a body of data, be it the data of coin-tossing, some natural process, or even Chinese sentences, requires the ability to compress that data.

[d] References[13,7] are suggested.

[e] MML is a Bayesian method of inductive and statistical inference and machine learning. MDL and MML are universally applicable to inference problems, such as problems of statistical parameter estimation[11,13,12] and problems of intrinsic classification[11], also known as unsupervised concept learning or mixture modelling. MML is also invariant under parameter transformation[13,4], and MDL and MML are guaranteed to converge with probability unity[13 1] to the correct inference. These methods are also efficient, converging as quickly as possible.

[f] Put another way, consider a variety of hypothesis, H, for explaining some data, D. By repeated application of Bayes's theorem, we have that $Pr(H|D) = Pr(H\&D)/Pr(D) = (1/Pr(D)) \times Pr(H) \times Pr(D|H)$. Since D and $1/Pr(D)$ can be assumed constant, maximising the posterior probability, $Pr(H|D)$, is equivalent to maximising $Pr(H)Pr(D|H)$, and to minimising the corresponding length of a two-part message, $-\log_2 Pr(H) - \log_2 Pr(D|H)$, for conveying an hypothesis, H, followed by D given H. This, the minimum message length (MML) principle, is an operational form of Ockham's razor since $-\log_2 Pr(H)$ concerns the (a priori) simplicity of the theory and $-\log_2 Pr(D|H)$ concerns how well the model fits the data, so minimising the message length gives us a simple hypothesis which fits the data well. The best compression gives the best theory and, indeed, the better the compression, the better the theory. In this sense, inductive learning equals (two-part) compression.

3 Physical limitations to Searle's Chinese Room

In practice, the Turing test will be carried out only over a finite number of steps, conservatively bounded above by a maximal human life span (e.g. 200 years) divided by (e.g.) a minimum acknowledged time period for humans to generate or recognise a syllable. For a conversation of fixed finite length, it seems plausible that a suitably large computer program could, in principle, be designed to pass this test by first exhaustively enumerating all of the finitely many nodes in this finite game tree and then prescribing a response in each case. Although this might initially seem plausible, consider the Chinese Room[8]. With an estimate of approximately 10^4 Chinese Mandarin characters with at least 10^3 in common usage, we conservatively estimate at least $(10^3)^5 = 10^{15}$ sentences of five characters or more which could possibly be exchanged in Chinese conversation after initial social pleasantries[g]. Our look-up table would thus need at least 10^{15} entries so that a response could be made to the first non-trivial part of the conversation. Moreover, being able to continue making sensible responses in a conversation of reasonable length will certainly require a look-up table with more entries than the currently estimated[h] number of elementary particles in the universe (approximately 10^{83}) if the universe is finite[i]. This contrasts rather starkly with the ability of humans to speak at least one language and to do much more using only an estimated 10^{12} or so neural processors[j]. And it means that we would be literally unable to write the look-up table in this universe (based on current theory, if the universe is finite) – even in principle.

Consider also the task of passing the Turing Test in the Chinese Room with an uncompressed look-up table in an infinite universe. We could conceivably store an arbitrarily large look-up table – one that could be used to simulate an hour or more or so, say, of conversation, even if this required the table to extend to distant galaxies. Assuming a fixed finite limit to the speed of transmission of information[k], if the Turing Test conversation is required to continue for long

[g] The fact that many sequences of characters will not form sensible sentences suggests that one should lower the estimate; on the other hand, the fact that sensible sentences can have many more than five characters more than compensates.

[h] we are grateful to Kurt Liffman for showing us calculations of how to use the critical particle density threshold to derive a figure closely approximating this oft-stated result.

[i] Note firstly that $(10^{15})^6 = 10^{90} > 10^{83}$. So if all sequences of input sentences were possible, only six consecutive inputs into the conversation would be needed. Perhaps certain sequences are ruled out (for example, if they contain gross non-sequiturs); but again, this is more than compensated for by the fact that conversations can last far longer than six exchanges.

[j] We are grateful to Joanne Luciano for directing us to a relevant reference.

[k] such as c, the speed of light from Einstein's theory of special relativity.

enough relative to (the cube of) this limiting speed, then the look-up table will need to be so large that, eventually, the response from the table's more distant entries will take a suspiciously long time to be given.

4 Intelligence, I.Q. and learning

So far, we have argued that inductive learning is compression and have pointed out that, without sufficient compression, Searle's Chinese Room eventually becomes very limited. In acknowledging the necessity of Turing's conditions for intelligence[10], it seems evident from I.Q. tests (e.g. [5]) that humans regard pattern recognition and inductive learning (to the best[l]) explanation as also being at least indicative of intelligence[m]. So, we would like to extend the test for intelligence ot require not just Turing's conditions, but also to require the ability to inductively learn (and hence to compress). We do not claim that our new criteria are *sufficient* for intelligence, but rather that they extend Turing's criteria while remaining *necessary* for intelligence.
One way of imposing our additional requirement on Turing's Test is to insist that the agent being subjected to the Turing Test not only pass the test but also have a concise, compressed representation of the subject domain. We do this because the Turing Test is a finite statistical test: we believe that a compressed method obtained by learning will be more likely to deal with likely with future questions in a reasonable amount of time than (e.g.) the brute-force rote-learnt Chinese room of Section 1.

5 Further questions – a statistical test and test administration

As above, the Turing Test is a finite statistical hypothesis test. Although the MML theory gives the best two-part compression and is the most probable theory and also gives both good "right" / "wrong" and probabilistic predictions, it is not necessarily the optimal "right" / "wrong" or probabilistic predictor. We therefore ask the following question:
Given two programs H_1 and H_2 respectively of lengths l_1 and l_2, $l_1 < l_2$, if H_1 and H_2 perform equally well on a Turing Test (or if $Pr(\text{Data}|\text{H}_1) = \Pr(\text{Data}|\text{H}_2)$), which, if either, should be predictively preferred for the future?

We have mentioned many human traits which seem to be hallmarks of "intel-

[l] or, as we would argue from Bayesian and information-theoretic grounds, MML.

[m] Some tests for intelligence also test for memory - or rote learning, and some for deductive learning. Rote learning is, of course, necessary to both store data and store the inference after compression. Deductive learning is, of course, necessary to combine inferences.

ligence": passing the Turing Test, inductive learning (and compression), rote learning (and memory) and deductive learning. But another things humans use their "intelligence" to do is to test others for intelligence, by Turing Tests, I.Q. tests or whatever. So, it seems reasonable to require that a sufficiently intelligent agent be able to *administer* the Turing Test (or, for that matter, an I.Q. test). Indeed, this idea can be iterated recursively.

References

1. A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.
2. G.J. Chaitin. On the length of programs for computing finite sequences. *Journal of the Association for Computing Machinery*, 13:547–549, 1966.
3. D.L. Dowe and A.R. Hájek. A computational extension to the Turing Test. Technical Report 97/322, Dept. of Computer Science, Monash University, Clayton 3168, Australia, September 1997.
4. D.L. Dowe and C.S. Wallace. Strict MML and Kolmogorov Complexity. to appear.
5. H.J. Eysenck. *Know your own I.Q.* Penguin, Harmondsworth, Middlesex, U.K., 1962.
6. Ming Li and P.M.B. Vitányi. *An Introduction to Kolmogorov Complexity and its applications.* Springer Verlag, New York, 1997.
7. J. J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore, 1989.
8. J.R. Searle. Minds, brains and programs. *Behavioural and Brain Sciences*, 3:417–457, 1980.
9. R.J. Solomonoff. A formal theory of inductive inference. *Information and Control*, 7:1–22,224–254, 1964.
10. A.M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.
11. C.S. Wallace and D.M. Boulton. An information measure for classification. *Computer Journal*, 11:185–194, 1968.
12. C.S. Wallace and D.L. Dowe. MML estimation of the von Mises concentration parameter. Tech rept TR 93/193, Dept. of Comp. Sci., Monash Univ., Clayton 3168, Australia, 1993. prov. accepted, Aust. J. Stat.
13. C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *J. Royal Statistical Society (Series B)*, 49:240–252, 1987.
14. J.G. Wolff. Learning and reasoning as information compression by multiple alignment, unification and search. In A. Gammerman, editor, *Computational Learning and Probabilistic Reasoning.* Wiley, New York, 1995.