# A Computational Extension to the Turing Test

David L. Dowe and Alan R. Hájek

Department of Computer Science, Monash University,
Clayton, Vic. 3168, Australia
HSS, California Institute of Technology, Pasadena,
California 91125, U.S.A.
e-mail: {dld@cs.monash.edu.au, ahajek@hss.caltech.edu}

August 17, 1997

### 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 and the problem of other minds.

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, Philosophy of AI, compression, Bayesian and Statistical Learning Methods, Machine Learning, Cognitive Modelling.

# 1 Introduction - the Turing Test and Chinese Room

'Intelligence', 'understanding' and 'learning' are multifarious, vague notions, not every aspect of which we can claim to catch in one fell swoop. Moreover, as Wittgenstein taught us, giving putatively necessary and sufficient conditions for some concept of ours can be a dangerous business. So, rather than courting such danger, we claim rather to capture a central notion of these cognitive concepts. Set aside rote learning (as occurs, for example, when one memorizes a list of the world's capital cities) and many aspects of deductive learning, and let us focus instead on inductive learning. We believe that this form of learning consists in the ability to compress data.

Turing introduced his famous test, "the Turing Test"[16], of (artificial) intelligence by proposing that the agent be tested for the ability to simulate by tele-type the conversational actions of a human. 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[13] gives a parallel example in which, instead of a machine trying to simulate humanness, a human endeavours to simulate an operational understanding of Chinese[1]. 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[2] 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 paper now proceeds by arguing that inductive learning amounts to compression, and that, without such compression, Searle's Chinese Room becomes very limited for a sufficiently long test. We then look at the issues of mindedness, a compression that one learns oneself vs. a compression that one is told of, whether a shorter or a longer program is to be preferred when both programs yield the same predictive accuracy to date, and

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

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

the issues of passing the Turing Test and administering the Turing Test.

# 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[23] 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[14, 2, 19] and has been successfully implemented in Minimum Message Length (MML)[19, 22] and Minimum Description Length (MDL)[12] applications ever since, both of which are related to Kolmogorov complexity[17, 4, 8].

For the reader possibly unfamiliar[3] with MML and MDL[4], 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: the data can now be summarized by recording just the $x_1$ values, and this functional relationship, and some error terms. More than that, we feel that 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 [5].

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. Now, understanding admits of degree: since you are reading this paper, you presumably understand English well; you may understand a certain foreign language moderately well, though not as well; and there are perhaps many foreign languages that you do not understand much at all. Correspondingly, you have implicitly or

[3] References[22, 15, 12] are suggested.

[4] 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[19, 22, 20, 18, 21] and problems of intrinsic classification[19, 21], also known as unsupervised concept learning or mixture modelling. MML is also invariant under parameter transformation[22, 21, 4], and MDL and MML are guaranteed to converge with probability unity [22, p241][18, 1] to the correct inference. These methods are also efficient, converging as quickly as possible.

[5] 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.

explicitly compressed the 'data' of English (its vocabulary and grammar) a great deal, and that certain foreign language somewhat (though less so). And you have not compressed the data of the foreign languages that you do not understand at all: indeed, they appear essentially 'random', (almost) unpatterned to you. Ultimate understanding, then, would appear to involve the ability to compress such data to the ultimate extent: that is, to know a minimum message length description of the data.

## 2.1 MML, compression, "laws" of nature and understanding

We can generalize the above still further. The universe apparently contains certain patterns; if these patterns are pervasive enough, they are good candidates for being "laws" of nature. Let us follow in the spirit of Mill[9], Ramsey[11] and Lewis[7], and regard the "laws" of nature as those (inferred, hypothesised) regularities that figure in the minimum message length[6] description of the universe. It is plausible that *understanding* the universe involves knowing its laws (much as understanding a language involves knowing its 'laws', that is, its rules); and that, given the data available about the universe, the best understanding of the universe would consist in the optimal (MML) inference learnt from this data.

Appealing to Bayesian, information-theoretic and MML notions, we argue above that inductive learning and (two-part)[7] compression are identical (or, at worst, very similar). Turing's original test does not require that the agent have a compressed representation of its knowledge, something which we argue in Section 3 can lead to problems. We also note that, other than by Turing's Test, intelligence is also measured by I.Q. (intelligence quotient), yet such tests (see, e.g., [5]) are very much concerned with problems of (optimal) pattern recognition and (optimal) inductive inference.

As further concepts are developed and inter-relations made between them, so the area of study is further compressed and so the agent comes to learn and better understand.

# 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[13]. 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

[6] or most economical

[7] The differences between two-part compression ($H$ and then $D$ given $H$) and one-part compression amount to the differences between MML[19, 22] and MDL[12], and are very small. Inductive learning is equivalent to two-part compression; and (see, e.g., [3]) prediction and one-part compression are equivalent.

initial social pleasantries[8]. 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[9] number of elementary particles in the universe (approximately $10^{83}$) if the universe is finite[10]. 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[11]. 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[12], if the Turing Test conversation is required to continue for long 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 Other minds, 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[16], it seems evident from I.Q. tests (e.g. [5]) that humans regard pattern recognition and inductive learning (to the best[13]) explanation as also being at least indicative of intelligence. (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.) 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

[8] 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.

[9] we are grateful to Kurt Liffman for showing us calculations of how to use the critical particle density[6] threshold to derive a figure closely approximating this oft-stated result.

[10] 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.

[11] We are grateful to Joanne Luciano for directing us to a relevant reference[10].

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

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

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 and we also suspect that a compressed method will be more reliable on future questions than (e.g.) the brute-force rote-learnt Chinese room of Section 1. A third reason that we desire compression (where possible) is that it is evidence of learning - if the agent is "intelligent" (and able to inductively learn), we would like it compress (where possible) the subject domain.

And, regarding minds, why do we believe of each other that we have minds? On one (rather bad and unreasonable) extreme, I can hypothesise you to have rote-learnt a Searle-style look-up table of English conversation; and on the other (rather good and reasonable) extreme, I can hypothesise you to have compressed and learnt much about English so that your $10^{12}$ or so neural processors (made up from your less than $10^{83}$ atoms) are sufficient for you to carry out a conversation. My belief that you can learn and compress and have done so in the past is part of why I attribute mindedness to you. If such considerations are legitimate when attributing intelligence to humans, then they presumably ought to apply equally well when attributing intelligence to machines – whatever "intelligence" is.

# 5 Further questions

## 5.1 A statistical test, and predictive reliability

The Turing Test is a finite statistical hypothesis test in that the test administrator collects finitely much (conversational or other) data from the agent and then hypothesises as to the intelligence or otherwise of the agent and, implicitly, how well the agent will perform on future inputs. Related to this, statistics and machine learning have notions of "right" / "wrong" predictions (which are rewarded as correct and incorrect) and probabilistic predictions (where probabilities are rewarded by their logarithms)[3]. 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 $Pr(\mathrm{Data}|\mathrm{H}_1) = \mathrm{Pr}(\mathrm{Data}|\mathrm{H}_2)$),
which, if either, should be predictively preferred for the future?
It is possible that a theorem [8, p340, Theorem 5.4.1] about PAC learning is relevant here.

## 5.2 Administration of the Turing Test

We have mentioned many human traits which seem to be hallmarks of "intelligence": 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 can be iterated recursively, with sufficiently intelligent agents being required to administer the administration of the Turing test. A rather loose analogy can be made along the lines that writing a paper (well) requires intelligence, so reviewing a paper well (administering the test well) requires intelligence and fulfilling editorial or Program Committee responsibilities (administering the test administration) well also requires intelligence.

# 6 Discussion and Conclusion

Inductive learning is tantamount to compression, I.Q. tests tend to be full of questions requiring such compression and pattern recognition, and our intuitive notions of intelligence, understanding and mindedness suggest that someone understands who has recognised and stored patterns (compressions) in some subject domain. And, just because an agent has passed a finite test, if they do not have an adequate compression, as with the limited capacity Chinese Room, the agent will not pass the test indefinitely in the future. We therefore argue that Turing's original test and our compression requirement constitute *necessary* conditions for intelligence. Intelligence requires the ability to compress; inductive learning is an act of such compression; and the state arrived at after such an act is one of increased understanding. Rote learning and deductive learning are also needed to store data and inferences and to combine inferences.

Let us now return to Searle's Chinese room not with rote-learnt responses, but rather with a rote-learning of someone else's compression (or algorithm). Let us also imagine a very mathematically adept human with a very good memory who is told to search a game tree and seek the path which will maximise a given mathematical function, closely mimicking the style of a human chess grandmaster. Each of these is a case of an agent who has a compression that she has not learnt (via compression) herself, an agent who possibly knows little or nothing about the rules of Chinese grammar and nothing about the rules of chess, and, indeed, to whom the strings “Qb6” and “Q–QN3” might be meaningless. In this sense, our enhanced test seems necessary but possibly not sufficient for intelligence.

# 7 Acknowledgments Section

The authors acknowledge helpful feedback from Chris Wallace, Ray Solomonoff, Graham Farr, Alex Byrne, Lloyd Allison, Kevin Korb and Paul Vitányi. This work was supported by Australian Research Council (ARC) Large Grants Nos. A49330656 and A49602504.

# 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, G.E. Farr, A.J. Hurst, and K.L. Lentin. Information-theoretic football tipping. In N. de Mestre, editor, *Proceedings of the Third Conference on Mathematics and Computers in Sport*, pages 233–241, Bond University, Qld., Australia, 1996.

[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] K.R. Lang. *Astrophysical formulae : a compendium for the physicist and astrophysicist.* Springer, 1974.

[7] D.K. Lewis. *Counterfactuals.* Harvard University Press, 1973.

[8] Ming Li and P.M.B. Vitányi. *An Introduction to Kolmogorov Complexity and its applications.* Springer Verlag, New York, 1997.

[9] J.S. Mill. *A System of Logic.* Parker, London, 1843.

[10] W.J.H. Naute and M. Feirtag. *Fundamental Neuroanatomy.* W.H. Freeman, 1986.

[11] F.P. Ramsey. *Universals of Law and of Fact.* Routledge and Kegan Paul, London, 1978.

[12] J. J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore, 1989.

[13] J.R. Searle. Minds, brains and programs. *Behavioural and Brain Sciences*, 3:417–457, 1980.

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

[15] R.J. Solomonoff. The discovery of algorithmic probability: A guide for the programming of true creativity. In P. Vitanyi, editor, *Computational Learning Theory: EuroCOLT'95*, pages 1–22. Springer-Verlag, 1995.

[16] A.M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.

[17] P.M.B. Vitányi and Ming Li. Ideal MDL and Its Relation to Bayesianism. In D.L. Dowe, K.B. Korb, and J.J. Oliver, editors, *Proc. Information, Statistics and Induction in Science (ISIS) Conference*, pages 282–291, Melbourne, 1996. World Scientific.

[18] C.S. Wallace. False Oracles and SMML Estimators. In *Proc. Information, Statistics and Induction in Science conference (ISIS'96)*, pages 304–316, Singapore, 1996. World Scientific. Was Tech Rept TR 89/128, Monash University, Australia, 1989.

[19] C.S. Wallace and D.M. Boulton. An information measure for classification. *Computer Journal*, 11:185–194, 1968.

[20] C.S. Wallace and D.L. Dowe. MML estimation of the von Mises concentration parameter. Technical report TR 93/193, Dept. of Comp. Sci., Monash Univ., Clayton, Vic. 3168, Australia, 1993. prov. accepted, Aust. J. Stat.

[21] C.S. Wallace and D.L. Dowe. Intrinsic classification by MML – the Snob program. In C. Zhang and et al., editors, *Proc. 7th Australian Joint Conf. on Artif. Intelligence*, pages 37–44. World Scientific, Singapore, 1994.

[22] C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society (Series B)*, 49:240–252, 1987.

[23] 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.