Tests of Machine Intelligence

Article · January 2008

DOI: 10.1007/978-3-540-77296-5_22 · Source: arXiv

Tests of Machine Intelligence

Shane Legg

IDSIA, Galleria 2, Manno-Lugano CH-6928, Switzerland

shane@vetta.org www.vetta.org/shane

Marcus Hutter

RSISE @ ANU and SML @ NICTA, Canberra, ACT, 0200, Australia

marcus@hutter1.net www.hutter1.net

December 2007

Abstract

Although the definition and measurement of intelligence is clearly of fundamental importance to the field of artificial intelligence, no general survey of definitions and tests of machine intelligence exists. Indeed few researchers are even aware of alternatives to the Turing test and its many derivatives. In this paper we fill this gap by providing a short survey of the many tests of machine intelligence that have been proposed.

Contents

1 Introduction

Despite solid progress on many fronts over the last 50 years, artificial intelligence is still a very young field with many of its greatest achievements, and some of its most fundamental problems, yet to be tackled. From a theoretical perspective, one of the most fundamental problems in the field is that the very concept of intelligence remains rather murky. This is somewhat true in the context of humans, but it is especially true when we consider machines which may have completely different sensors, bodies, cognitive capacities and live in different environments to ourselves. What does "intelligence" mean for a machine? Perhaps the first attempt to answer this question, and certainly the only attempt that most researchers are aware of, is Alan Turing's famous imitation game [33]. Turing recognised how difficult it would be to explicitly define intelligence and thus attempted to sidestep the issue completely. Although this was a clever move, it leaves us with a test of machine intelligence that tells us almost nothing about what intelligence actually is, and thus is of little use as a foundation, either theoretical or practical, for our research.

Since then, a few bold researchers have tried to tackle this difficult problem in a more satisfactory way by proposing various definitions and tests of machine intelligence. By and large, these proposals have been ignored by the community. Indeed to the best of our knowledge, no general survey of tests and definitions of intelligence for machines has ever been published.

We feel that to ignore a question as fundamental as the definition of machine intelligence is a serious mistake. In any science, issues surrounding fundamental definitions and methods of measurement play a central role and form the foundation on which theoretical advances are constructed and practical advances are measured. If we are to truly advance as a field over the next 50 years, we will need to return to this most central of problems in order to secure what artificial intelligence is and what it aims for. As a first step in this direction, it is necessary that researchers are at least aware of the many alternatives to Turing's tests that have been proposed. In this paper we hope to partly meet this need by providing the first general survey of tests and definitions of machine intelligence.

2 Turing test and derivatives

The classic approach to determining whether a machine is intelligent is the so called Turing test [33] which has been extensively debated over the last 50 years [26]. Turing realised how difficult it would be to directly definite intelligence and thus attempted to side step the issue by setting up his now famous imitation game: If human judges cannot effectively discriminate between a computer and a human through teletyped conversation, then we must conclude that the computer is intelligent.

Though simple and clever, the test has attracted much criticism. Block and Searle argue that passing the test is not *sufficient* to establish intelligence [3, 28, 7].

Essentially they both argue that a machine could appear to be intelligent without having any "real intelligence", perhaps by using a very large table of answers to questions. While such a machine might be impossible in practice due to the vast size of the table required, it is not logically impossible. In which case an unintelligent machine could, at least in theory, consistently pass the Turing test. Some consider this to bring the validity of the test into question. In response to these challenges, even more demanding versions of the Turing test have been proposed such as the Total Turing test [11], the Truly Total Turing test [27] and the inverted Turing test [35]. Dowe argues that the Turing test should be extended by ensuring that the agent has a compressed representation of the domain area, thus ruling out look-up table counter arguments [6]. Of course these attacks on the Turing test can be applied to any test of intelligence that considers only a system's external behaviour, that is, most intelligence tests.

A more common criticism is that passing the Turing test is not *necessary* to establish intelligence. Usually this argument is based on the fact that the test requires a machine to have a highly detailed model of human knowledge and patterns of thought, making it a test of humanness rather than intelligence [9, 8]. Indeed even small things like pretending to be *unable* to perform complex arithmetic quickly and faking human typing errors become important, something which clearly goes against the purpose of the test.

The Turing test has other problems as well. Current AI systems are a long way from being able to pass an unrestricted Turing test. From a practical point of view this means that the full Turing test is unable to offer much guidance to our work. Indeed, even though the Turing test is the most famous test of machine intelligence, almost no current research in artificial intelligence is specifically directed toward being able to pass it. Unfortunately, simply restricting the domain of conversation in the Turing test to make the test easier, as is done in the Loebner competition [22], is not sufficient. With restricted conversation possibilities the most successful Loebner entrants are even more focused on faking human fallibility, rather than anything resembling intelligence [15]. Perhaps a better alternative then is to test whether a machine can imitate a child (see for example the tests described in Sections 4 and 5). Finally, the Turing test returns different results depending on who the human judges are. Its unreliability has in some cases lead to clearly unintelligent machines being classified as human, and at least one instance of a human actually failing a Turing test. When queried about the latter, one of the judges explained that "no human being would have that amount of knowledge about Shakespeare"[29].

3 Compression tests

Mahoney has proposed a particularly simple solution to the binary pass or fail problem with the Turing test: Replace the Turing test with a text compression test [23]. In essence this is somewhat similar to a "Cloze test" where an individual's com-

prehension and knowledge in a domain is estimated by having them guess missing words from a passage of text.

While simple text compression can be performed with symbol frequencies, the resulting compression is relatively poor. By using more complex models that capture higher level features such as aspects of grammar, the best compressors are able to compress text to about 1.5 bits per character for English. However humans, which can also make use of general world knowledge, the logical structure of the argument etc., are able to reduce this down to about 1 bit per character. Thus the compression statistic provides an easily computed measure of how complete a machine's model of language, reasoning and domain knowledge are, relative to a human.

To see the connection to the Turing test, consider a compression test based on a very large corpus of dialogue. If a compressor could perform extremely well on such a test, this is mathematically equivalent to being able to determine which sentences are probable at a given point in a dialogue, and which are not (for the equivalence of compression and prediction see [2]). Thus, as failing a Turing test occurs when a machine (or person!) generates a sentence which would be improbable for a human, extremely good performance on dialogue compression implies the ability to pass a Turing test.

A recent development in this area is the Hutter Prize [17]. In this test the corpus is a 100 MB extract from Wikipedia. The idea is that this should represent a reasonable sample of world knowledge and thus any compressor that can perform very well on this test must have a good model of not just English, but also world knowledge in general.

One criticism of compression tests is that it is not clear whether a powerful compressor would easily translate into a general purpose artificial intelligence.

4 Linguistic complexity

A more linguistic approach is taken by the HAL project at the company Artificial Intelligence NV [32]. They propose to measure a system's level of conversational ability by using techniques developed to measure the linguistic ability of children. These methods examine things such as vocabulary size, length of utterances, response types, syntactic complexity and so on. This would allow systems to be "...assigned an age or a maturity level beside their binary Turing test assessment of 'intelligent' or 'not intelligent' "[31]. As they consider communication to be the basis of intelligence, and the Turing test to be a valid test of machine intelligence, in their view the best way to develop intelligence is to retrace the way in which human linguistic development occurs. Although they do not explicitly refer to their linguistic measure as a test of intelligence, because it measures progress towards what they consider to be a valid intelligence test, it acts as one.

5 Multiple cognitive abilities

A broader developmental approach is being taken by IBM's Joshua Blue project [1]. In this project they measure the performance of their system by considering a broad range of linguistic, social, association and learning tests. Their goal is to first pass what they call a "toddler Turing test", that is, to develop an AI system that can pass as a young child in a similar setup to the Turing test. As yet, this test is not fully specified.

Another company pursuing a similar developmental approach based on measuring system performance through a broad range of cognitive tests is the a2i2 project at Adaptive AI [34]. Rather than toddler level intelligence, their current goal to is work toward a level of cognitive performance similar to that of a small mammal. The idea being that even a small mammal has many of the key cognitive abilities required for human level intelligence working together in an integrated way. While this might be useful to guide the development of moderate intelligence, it is unknown whether it will scale to higher levels of intelligence. The specific tests being used have not been published.

6 Competitive games

The Turing Ratio method of Masum et al. has more emphasis on tasks and games rather than cognitive tests. They propose that "...doing well at a broad range of tasks is an empirical definition of 'intelligence'."[24] To quantify this they seek to identify tasks that measure important abilities, admit a series of strategies that are qualitatively different, and are reproducible and relevant over an extended period of time. They suggest a system of measuring performance through pairwise comparisons between AI systems that is similar to that used to rate players in the international chess rating system. The key difficulty however, which the authors acknowledge is an open challenge, is to work out what these tasks should be, and to quantify just how broad, important and relevant each is. In our view these are some of the most central problems that must be solved when attempting to construct an intelligence test and thus this approach is incomplete in its current state.

7 Collection of psychometric tests

An approach called Psychometric AI tries to address the problem of what to test for in a pragmatic way. In the view of Bringsjord and Schimanski, "Some agent is intelligent if and only if it excels at all established, validated tests of [human] intelligence."[4] They later broaden this to also include "tests of artistic and literary creativity, mechanical ability, and so on." With this as their goal, their research is focused on building robots that can perform well on standard psychometric tests

designed for humans, such as the Wechsler Adult Intelligent Scale and Raven Progressive Matrices.

As effective as these tests are for humans, they seem inadequate for measuring machine intelligence as they are highly anthropocentric and embody basic assumptions about the test subject that are likely to be violated by computers. For example, consider the fundamental assumption that the test subject is not simply a collection of specialised algorithms designed only for answering common IQ test questions. While this is obviously true of a human, or even an ape, it may not be true of a computer. The computer could be nothing more than a collection of specific algorithms designed to identify patterns in shapes, predict number sequences, write poems on a given subject or solve verbal analogy problems — all things that AI researchers have worked on. Such a machine might be able to obtain a respectable IQ score [25], even though outside of these specific test problems it would be next to useless. If we try to correct for these limitations by expanding beyond standard tests, as Bringsjord and Schimanski seem to suggest, this once again opens up the difficulty of exactly what, and what not, to test for. Psychometric AI, at least as it is currently formulated, only partially addresses this central question.

8 Smith's test

The basic structure of Smith's test is that an agent faces a series of problems that are generated by an algorithm [30]. In each iteration the agent must try to produce the correct response to the problem that it has been given. The problem generator then responds with a score of how good the agent's answer was. If the agent so desires it can submit another answer to the same problem. At some point the agent requests to the problem generator to move onto the next problem and the score that the agent received for its last answer to the current problem is then added to its cumulative score. Each interaction cycle counts as one time step and the agent's intelligence is then its total cumulative score considered as a function of time. In order to keep things feasible, the problems must all be in P, i.e. the solution must be verifiable in polynomial time.

We have two main criticisms of Smith's definition. Firstly, while for practical reasons it might make sense to restrict problems to be in P, we do not see why this practical restriction should be a part of the very definition of intelligence as Smith suggests. If some breakthrough meant that agents could solve difficult problems in not just P but sometimes in NP as well, then surely these new agents would be more intelligent?

Secondly, while the definition is somewhat formally defined, it still leaves open the important question of what exactly the tests should be. Smith suggests that researchers should dream up tests and then contribute them to some common pool of tests. As such, this is not a fully specified test.

9 C-Test

One perspective among psychologists who support the g-factor view of intelligence, is that intelligence is "the ability to deal with complexity"[10]. Thus in a test of intelligence the most difficult questions are the ones that are the most complex because these will, by definition, require the most intelligence to solve. It follows then that if we could formally define and measure the complexity of test problems we could construct a formal test of intelligence. The possibility of doing this was perhaps first suggested by the complexity theorist Chaitin [5]. While this path requires numerous difficulties to be dealt with, we believe that it is the most natural and offers many advantages: It is formally motivated, precisely defined and potentially could be used to measure the performance of both computers and biological systems on the same scale without the problem of bias towards any particular species or culture.

One intelligence test that is based on formal complexity theory is the C-Test from Hernández [13, 14]. This test consists of a number of sequence prediction and abduction problems similar to those that appear in many standard IQ tests. Similar to standard IQ tests, the C-Test always ensures that each question has an unambiguous answer in the sense that there is always one hypothesis that is consistent with the observed pattern that has significantly lower complexity than the alternatives. The key difference to sequence problems that appear in standard intelligence tests is that the questions are based on a formally expressed measure of complexity, namely Levin's computable Kt complexity [20] (rather than Kolmogorov's incomputable complexity [21]) to get a practical test. In order to retain the invariance property of Kolmogorov complexity, Levin complexity requires the additional assumption that the universal Turing machines are able to simulate each other in linear time.

The test has been successfully applied to humans with intuitively reasonable results [14, 12]. As far as we know, this is the only formal definition of intelligence that has so far produced a usable test of intelligence.

One criticism of the C-Test and Smith's tests is that the way intelligence is measured is essentially static, that is, the environments are passive. We believe that dynamic testing in active environments is a better measure of a system's intelligence. To put this argument another way: Succeeding in the real world requires you to be more than an insightful spectator! One must carefully choose actions knowing that these may affect the future.

10 Universal intelligence

Another complexity based test is the *universal intelligence* test [19]. Unlike the C-Test and Smith's test, universal intelligence tests the performance of an agent in a fully interactive environment. This is done by using the reinforcement learning framework in which the agent sends its *actions* to the environment and receives *observations* and *rewards* back. The agent tries to maximise the amount of reward

it receives by learning about the structure of the environment and the goals it needs to accomplish in order to receive rewards.

Formally, the process of interaction produces an increasing history $o_1r_1a_1o_2r_2a_2o_3r_3a_3o_4\ldots$ of observations o, rewards $r \geq 0$, and actions a. The agent is simply a function, denoted by π, which is a probability measure over actions conditioned on the current history, for example, $\pi(a_3|o_1r_1a_1o_2r_2)$. The environment, denoted μ, is similarly defined: $\mu(o_kr_k|o_1r_1a_1o_2r_2a_2\ldots o_{k-1}r_{k-1}a_{k-1})$. The performance of agent π in environment μ can be measured by its total expected reward $V_\mu^\pi := \mathbf{E}[\sum_{i=1}^\infty r_i|\mu,\pi]$, called value. The largest interesting class of environments is the class E of all computable probability distributions μ. For technical reasons, the values are assumed to be bounded by some constant c.

To get a single performance measure V_μ^π is averaged over all $\mu \in E$. As there are an infinite number of environments, with no bound on their complexity, it is impossible to take the expected value with respect to a uniform distribution — some environments must be weighted more heavily than others. Considering the agent's perspective on the problem, it is the same as asking: Given several different hypotheses which are consistent with the observations, which hypothesis should be considered the most likely? This is a fundamental problem in inductive inference for which the standard solution is to invoke Occam's razor: *Given multiple hypotheses which are consistent with the data, simpler ones should be preferred.* As this is generally considered the most intelligent thing to do, one should test agents in such a way that they are, at least on average, rewarded for correctly applying Occam's razor. This means that the a priori distribution over environments should be weighted towards simpler environments.

As each environment μ is described by a computable measure, their complexity can be measured with Kolmogorov complexity $K(\mu)$, which is simply the length of the shortest program that computes μ [21]. The right a priori weight for μ is $2^{-K(\mu)}$. We can now define the *universal intelligence* of an agent π to simply be its expected performance,

$$\Upsilon(\pi) := \sum_{\mu\in E} 2^{-K(\mu)} V_\mu^\pi.$$

By construction, universal intelligence measures the general ability of an agent to perform well in a very wide range of environments, similar to the essence of many informal definitions of intelligence [18]. The definition places no restrictions on the internal workings of the agent; it only requires that the agent is capable of generating output and receiving input which includes a reward signal. If we wish to bias the test to reflect world knowledge then we can condition the complexity measure. For example, use $K(\mu|D)$ where D is some set of background knowledge such as Wikipedia.

By considering V_μ^π for a number of basic environments, such as small MDPs, and agents with simple but very general optimisation strategies, it is clear that Υ correctly orders the relative intelligence of these agents in a natural way. A very high value of Υ would imply that an agent is able to perform well in many

environments. The maximal agent with respect to Υ is the theoretical AIXI agent which has been shown to have many strong optimality properties [16]. These results confirm that agents with high universal intelligence are indeed very powerful and adaptable. Universal intelligence spans simple adaptive agents right up to super intelligent agents like AIXI. The test is completely formally specified in terms of fundamental concepts such as universal Turing computation and complexity and thus is not anthropocentric.

A test based on Υ would evaluate the performance of an agent on a large sample of simulated environments, and then combine the agent's performance in each environment into an overall intelligence value. The key challenge that needs to be dealt with is to find a suitable replacement for the incomputable Kolmogorov complexity function, possibly Levin's Kt complexity [20], as is done by the C-Test.

11 Summary

We end this survey with a comparison of the various tests considered. Table 1 rates each test according to the properties described below. Although we have attempted to be as fair as possible, some of the scores we give on this table will naturally be debatable. Nevertheless, we hope that it provides a rough overview of the relative strengths and weaknesses of the proposals.

Valid: A test of intelligence should capture intelligence and not some related quantity. *Informative*: The result should be a scalar value, or perhaps a vector. *Wide range*: A test should cover low levels of intelligence up to super intelligence. *General*: Ideally we would like to have a very general test that could be applied to everything from a fly to a machine learning algorithm. *Dynamic*: A test should directly take into account the ability to learn and adapt over time. *Unbiased*: A test should not be biased towards any particular culture, species, etc. *Fundamental*: We do not want a test that needs to be changed from time to time due to changing technology and knowledge. *Formal*: The test should be precisely defined, ideally using mathematics. *Objective*: The test should not appeal to subjective assessments such as the opinions of human judges. *Fully Defined*: Has the test been fully defined, or are parts still unspecified? *Universal*: Is the test universal, or is it anthropocentric? *Practical*: A test should be able to be performed quickly and automatically. *Test vs. Def*: Finally we note whether the proposal is more of a test, more of a definition, or something in between.

Acknowledgements.

This work was supported by the Swiss NSF grant 200020-107616.

Table 1: In the table ● means "yes", • means "debatable", · means "no", and ? means unknown. When something is rated as unknown that is usually because the test in question is not sufficiently specified.

Intelligence Test	Valid	Informative	Wide Range	General	Dynamic	Unbiased	Fundamental	Formal	Objective	Fully Defined	Universal	Practical	Test vs. Def.
Turing Test	•	·	·	·	●	·	·	·	·	●	·	●	T
Total Turing Test	•	·	·	·	●	·	·	·	·	●	·	·	T
Inverted Turing Test	•	•	·	·	●	·	·	·	·	●	·	●	T
Toddler Turing Test	•	·	·	·	●	·	·	·	·	·	·	●	T
Linguistic Complexity	•	●	•	·	·	·	·	●	●	·	●	●	T
Text Compression Test	•	●	●	•	·	•	•	●	●	●	•	●	T
Turing Ratio	•	●	●	●	?	?	?	?	?	·	?	?	T/D
Psychometric AI	●	●	•	●	?	•	·	•	•	•	·	•	T/D
Smith's Test	•	●	●	•	·	?	●	●	●	·	?	•	T/D
C-Test	•	●	●	•	·	●	●	●	●	●	●	●	T/D
Universal Intelligence	●	●	●	●	●	●	●	●	●	●	●	·	D

References

[1] N. Alvarado, S. Adams, S. Burbeck, and C. Latta. Beyond the Turing test: Performance metrics for evaluating a computer simulation of the human mind. In *Performance Metrics for Intelligent Systems Workshop*, Gaithersburg, MD, USA, 2002. North-Holland.

[2] T. C. Bell, J. G. Cleary, and I. H. Witten. *Text compression.* Prentice Hall, 1990.

[3] N. Block. Psychologism and behaviorism. *Philosophical Review*, 90:5–43, 1981.

[4] S. Bringsjord and B. Schimanski. What is artificial intelligence? Psychometric AI as an answer. *Eighteenth International Joint Conference on Artificial Intelligence*, 18:887–893, 2003.

[5] G. J. Chaitin. Gödel's theorem and information. *International Journal of Theoretical Physics*, 22:941–954, 1982.

[6] D. L. Dowe and A. R. Hajek. A non-behavioural, computational extension to the Turing test. In *International Conference on Computational Intelligence & Multimedia Applications (ICCIMA '98)*, pages 101–106, Gippsland, Australia, 1998.

[7] J. Eisner. Cognitive science and the search for intelligence. *Invited paper presented to the Socratic Society, University of Cape Town*, 1991.

[8] K. M. Ford and P. J. Hayes. On computational wings: Rethinking the goals of artificial intelligence. *Scientific American*, Special Edition(4), 1998.

[9] R. M. French. Subcognition and the limits of the Turing test. *Mind*, 99:53–65, 1990.

[10] L. S. Gottfredson. Why g matters: The complexity of everyday life. *Intelligence*, 24(1):79–132, 1997.

[11] S. Harnad. Minds, machines and Searle. *Journal of Theoretical and Experimental Artificial Intelligence*, 1:5–25, 1989.

[12] J. Hernández-Orallo. Beyond the Turing test. *Journal of Logic, Language and Information*, 9(4):447–466, 2000.

[13] J. Hernández-Orallo. On the computational measurement of intelligence factors. In *Performance Metrics for Intelligent Systems Workshop*, pages 1–8, Gaithersburg, MD, USA, 2000.

[14] J. Hernández-Orallo and N. Minaya-Collado. A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. In *Proceedings of the International Symposium of Engineering of Intelligent Systems (EIS'98)*, pages 146–163. ICSC Press, 1998.

[15] J. L. Hutchens. How to pass the Turing test by cheating. *www.cs.umbc.edu/471/current/papers/hutchens.pdf*, 1996.

[16] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005. 300 pages, http://www.hutter1.net/ai/uaibook.htm.

[17] M. Hutter. The Human knowledge compression prize. *http://prize.hutter1.net*, 2006.

[18] S. Legg and M. Hutter. A collection of definitions of intelligence. In B. Goertzel and P. Wang, editors, *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, volume 157 of *Frontiers in Artificial Intelligence and Applications*, pages 17–24, Amsterdam, NL, 2007. IOS Press.

[19] S. Legg and M. Hutter. A formal measure of machine intelligence. In *Proc. 15th Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn'06)*, pages 73–80, Ghent, 2006.

[20] L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9:265–266, 1973.

[21] M. Li and P. M. B. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 2nd edition, 1997.

[22] H. G. Loebner. The Loebner prize — The first Turing test. *http://www.loebner.net/Prizef/loebner-prize.html*, 1990.

[23] M. V. Mahoney. Text compression as a test for artificial intelligence. In *AAAI/IAAI*, 1999.

[24] H. Masum, S. Christensen, and F. Oppacher. The Turing ratio: Metrics for open-ended tasks. In *GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference*, pages 973–980, New York, 2002. Morgan Kaufmann Publishers.

[25] P. Sanghi and D. L. Dowe. A computer program capable of passing I.Q. tests. In *Proc. 4th ICCS International Conference on Cognitive Science (ICCS'03)*, pages 570–575, Sydney, NSW, Australia, 2003.

[26] A. Saygin, I. Cicekli, and V. Akman. Turing test: 50 years later. *Minds and Machines*, 10, 2000.

[27] P. Schweizer. The truly total Turing test. *Minds and Machines*, 8:263–272, 1998.

[28] J. Searle. Minds, brains, and programs. *Behavioral & Brain Sciences*, 3:417–458, 1980.

[29] S. Shieber. Lessons from a restricted Turing test. *CACM: Communications of the ACM*, 37, 1994.

[30] W. D. Smith. Mathematical definition of "intelligence" (and consequences). *http://math.temple.edu/~ wds/homepage/works.html*, 2006.

[31] A. Treister-Goren, J. Dunietz, and J. L. Hutchens. The developmental approach to evaluating artificial intelligence – a proposal. In *Performance Metrics for Intelligence Systems*, 2000.

[32] A. Treister-Goren and J. L. Hutchens. Creating AI: A unique interplay between the development of learning algorithms and their education. In *Proceeding of the First International Workshop on Epigenetic Robotics*, 2001.

[33] A. M. Turing. Computing machinery and intelligence. *Mind*, October 1950.

[34] P. Voss. Essentials of general intelligence: The direct path to AGI. In B. Goertzel and C. Pennachin, editors, *Artificial General Intelligence*. Springer-Verlag, 2005.

[35] S. Watt. Naive psychology and the inverted Turing test. *Psycoloquy*, 7(14), 1996.