*Diametros* 16 (2019), 59: 35–47
doi: 10.33392/diam.1125

# The Minimum Intelligent Signal Test (MIST) as an Alternative to the Turing Test

– Paweł Łupkowski, Patrycja Jurowska –

**Abstract.** The aim of this paper is to present and discuss the issue of the adequacy of the Minimum Intelligent Signal Test (MIST) as an alternative to the Turing Test. MIST has been proposed by Chris McKinstry as a better alternative to Turing's original idea. Two of the main claims about MIST are that (1) MIST questions exploit commonsense knowledge and as a result are expected to be easy to answer for human beings and difficult for computer programs; and that (2) the MIST design aims at eliminating the problem of the role of judges in the test. To discuss these design assumptions we will present Peter D. Turney's PMI-IR algorithm which allows for MIST-type questions to be answered. We will also present and discuss the results of our own study aimed at the judge problem for MIST.
**Keywords:** Turing test, Minimum Intelligent Signal Test, subcognitive questions, commonsense knowledge, Artificial Intelligence

Published online: 31 March 2019

## 1. Introduction

In his seminal paper Alan Turing proposed a test for machines.[1] A machine would pass the test if it were capable of having a convincing, human-like tele-typed conversation with a human judge (the parties to the test cannot see or hear one another). Since then, the Turing test (hereafter TT) has been widely discussed by philosophers, psychologists, computer scientists and cognitive scientists.[2] Despite the fact that it was proposed

Paweł Łupkowski
Adam Mickiewicz University, Poznań
Institute of Psychology
Department of Logic and Cognitive Science
Reasoning Research Group
Szamarzewskiego 89a
60-568 Poznań
e-mail: pawel.lupkowski@amu.edu.pl
Patrycja Jurowska,
Adam Mickiewicz University, Poznań
Institute of Psychology
Reasoning Research Group
Szamarzewskiego 89a
60-568 Poznań
e-mail: pjurowska@gmail.com

[1] Turing (1950).
[2] See e.g., Konar (2000); Harnish (2002).

more than sixty years ago, TT is still considered as a fruitful theoretical idea.[3] It is worth stressing that the TT idea has also practical applications — see e.g., the Loebner contest[4] or CAPTCHA systems.[5]

The judge's perspective in TT is one of the central issues when we try to evaluate this test setting[6] and had already been noticed by Turing. His suggestion was that the interrogator should be a person who is not an expert in the field of computing machines.[7] Such a requirement stemmed from the fact that Turing was aware that the beliefs and knowledge of the interrogator might play an important role in the running of the test. The judge bias is often pointed out as one of the main drawbacks of the Turing test. Ned Block, for example, writes:

> [c]onstrued as a proposal about how to make the concept of intelligence precise, there is a gap in Turing's proposal: we are not told how the judge is to be chosen. A judge who was a leading authority on genuinely intelligent machines might know how to tell them apart from people. For example, the expert may know that current intelligent machines get certain problems right that people get wrong. [...] A stupid judge, or one who has had no contact with technology, might think that a radio was intelligent. People who are naive about computers are amazingly easy to fool [...].[8]

This issue has a very practical dimension, as the problem of selecting judges for TT becomes even more important when we think of the Loebner contest (hereafter LC). LC to a large extent may be treated as a practical realization of TT and, as such, it reveals certain problems with TT's design. The analysis of transcripts of 2009–2012 Loebner contest editions sheds more light on the role of the judge in LC: "The biggest drawback of LC is that the judge knows that the conversation is taking place with a human and a program, and the task is only to decide which is which. That makes it a much harder task for the program. It is not enough to exhibit intelligent behaviors and hold a decent conversation — the program has to be at least as human-like as the competing human."[9]

What is more, one may argue that judges will never have a "normal" conversation in LC. The reason for this is that they are placed in the test-like environment with the main aim set at identifying contestants.

Several solutions to the judge bias problem may be found in the literature. They range from Loebner's idea[10] to employ journalists as judges to the concept of introducing a kind of protocol for the Loebner contest (regulating the range of problems and questions allowed for the contest).[11] However, there are two concepts put forward to modify the general TT setting which we find especially interesting and promising. One

[3] See Saygin et al. (2001); Shieber (2004); Epstein et al. (2009) or Łupkowski and Wiśniewski (2011).
[4] Loebner (2009).
[5] Ahn et al. (2003).
[6] See the discussion in Łupkowski (2010) and (2011).
[7] Cf. Turing (1950): 442; Newman et al. (1952): 4.
[8] Block (1995): 379.
[9] Łupkowski and Rybacka (2016): 361.
[10] Loebner (2009).
[11] See Garner (2009) or Watt (2009).

of them is the *Unsuspecting Turing Test* (UTT) and the other one is the *Minimum Intelligent Signal Test* (MIST). Both proposals draw on Turing's original idea, but change the way the testing is performed, and as such they aim at eliminating the judge's issue from the picture. What is also appealing, UTT and MIST are designed in such a way that it makes possible to test their main assumptions.

The Unsuspecting Turing Test was proposed in 1994 a short paper by Michael Mauldin.[12] Mauldin used the *TinyMud* game (a text-based multiplayer RPG game) and introduced a bot (named *ChatterBot*) into the game. He observed that the bot was often taken for a human player. As Mauldin writes: "The ChatterBot succeeds in the TinyMud world because it is an *unsuspecting Turing test*, meaning that the players assume everyone else playing is a person, and will give the ChatterBot the benefit of the doubt until it makes a major gaffe."[13]

The Minimum Intelligent Signal Test was proposed by Chris McKinstry.[14] His idea is to set such rules for TT that will allow to perform it automatically. McKinstry claims that it would be possible if only yes/no questions would be allowed in the test and if an interrogator would evaluate patterns of answers instead of single answers. The very idea is to compare patterns of answers to the same set of questions obtained from a machine with the ones obtained from human beings.

In this paper we focus our attention on MIST with our aim being an evaluation of it as an alternative to TT. We start by introducing MIST design and its core assumptions: (1) that MIST questions should be easy for humans and difficult for machines and (2) that MIST results should be easy to evaluate. This is followed by discussing these assumptions with respect to the PMI-IR algorithm proposed by Peter D. Turney, which was designed to answer MIST-like questions. Then we present the results of our own study aimed at the judge problem for MIST.

## 2. The Minimum Intelligent Signal Test

MIST was first described by McKinstry in a very short (two-page) paper "Minimum Intelligence Signal Test: an Objective Turing Test."[15] McKinstry claims that the main problem with the TT setting is that it gives us a binary answer[16] when it comes to machines' intelligence:

[12] See Mauldin (1994) and discussion in Mauldin (2009).

[13] Mauldin (1994): 17. At this point it may be noticed that TT and its alternatives (like UTT and MIST mentioned here) put stress on the artificial agent's performance. As it was pointed out by an anonymous referee it would be beneficial to consider the ability to make mistakes itself as the criterion of mentality. This idea is explored e.g., within the theory of minds as semiotic systems — see Fetzer (1995), (1997).

[14] McKinstry (1997), (2009).

[15] McKinstry (1997). More extensive description of MIST may be found in McKinstry (2009).

[16] It is worth mentioning that Turing's idea is not that simplistic. He assumed that an agent should be tested long enough to gain more reliable results. As it is clearly stated in "Can Digital Computers Think": "We had better suppose that each jury has to judge quite a number of times, and that sometimes they really are dealing with a man and not a machine. That will prevent them saying 'It must be a machine' every time without proper consideration". Newman et al. (1952): 5; see also Turing (1950): 442.

> The 'all-or-nothing' nature of the Turing Test makes it of no use in the creation or measurement of emerging intelligent systems — it can only tell us if we have an intelligent system after the fact. What we really need is a Turing-like test that admits degrees and treats intelligence as at least a human continuum — a test that would allow us to measure the minimum amounts of global human intelligence that are the precursors of full adult human intelligence — a test that can be easily automated so it can be executed at machine speeds.[17]

To achieve such a goal, McKinstry proposes a test in which only yes/no questions are allowed. This ensures that the tested agent will not have the opportunity to provide misleading or evasive answers (as it is often visible in LC transcripts) — it has to provide a simple "yes" or "no" to a given question.[18] Such a setting allows for automatization with respect to running the test and also for evaluating provided answers. This — in theory — should eliminate judge's bias. McKinstry claims that evaluation of MIST boils down to a simple comparison of answers provided by a tested agent with those provided to the same questions by human participants.

As for the content of questions in MIST McKinstry writes that they should address our commonsense knowledge about the world, as e.g., "Do you exist?", "Are you a rock?", "Are you a human being?".[19] He also claims that the subcognitive questions proposed by Robert French would be a perfect inspiration for MIST questions. As French puts it:

> Surely, we would not want to limit a Turing Test to questions like 'What is the capital of France?' or 'How many sides does a triangle have?.' If we admit that intelligence in general must have something to do with categorization, analogy making, and so on, we will of course want to ask questions that test these capacities. But these are the very questions that will allow us, unfailingly, to unmask the computer.[20]

Subcognitive questions should be designed to reveal low-level cognitive structures, that is "the subconscious associative network in human minds that consists of highly overlapping activatable representations of experience."[21] This assumption makes such questions difficult for machines, as they require acquiring intelligence about the world by experiencing it in the way human beings do during their lifetime. Examples of such questions are the following:

- On a scale of 0 (completely implausible) to 10 (completely plausible), please rate 'Flugly' as the name a child might give its favorite teddy bear.
- On a scale of 0 (completely implausible) to 10 (completely plausible), please rate banana splits as medicine.

[17] McKinstry (2009): 286.
[18] McKinstry (2009): 289.
[19] See McKinstry (2009): 290.
[20] French (1990): 63.
[21] French (1990): 56–57.

- On a scale of 0 (completely implausible) to 10 (completely plausible), please rate purses as weapons.
- Please rate the following smells (1 — very bad, 10 — very nice):
  a) Newly cut grass;
  b) Freshly baked bread;
  c) A wet bath towel;
  d) Ground pepper.[22]

McKinstry's idea was to create a database of questions of this kind, where each question is connected with an answer. Such a pair is called a *mindpixel* (see examples of such question-answer pairs presented in the Appendix of this paper). In order to collect such data McKinstry started the MindPixel project, for which internet users could contribute *mindpixels* to a large database (the project was active from 2000 to 2005). As McKinstry put it in the online interview: "The first phase is a completely public, internet based effort. All the data it will be collecting will come from average people, with no specific training in AI or psychology."[23] Such a corpus will be then used for MIST.

As for the MIST procedure, McKinstry describes it in the following manner.[24]

1. N items (i.e. yes/no questions) are generated. For all these items, humans should be able to provide an answer (affirmative or negative). The distribution of items should be that for about 50% expected reaction should be positive and negative for the rest (this proportion is aimed at reducing the bias for answering yes/no questions.[25] At this stage, we also collect the answers from human participants and as an effect we obtain a large corpus of questions and human-intelligence answers.
2. Items are presented, and responses recorded. Items should be presented in a random order and on subsequent re-trials, item order is re-randomized.
3. For each item a judge evaluates an item/response pair as either consistent or inconsistent with human intelligence. McKinstry claims that this grading procedure may be easily automated, reducing the chance of the grading error or an unforeseen bias.
4. Generate Score. The result is not "all or nothing" for a tested machine. We only obtain the percentage in which the machine's answers are evaluated as human-like intelligent. This level should be more than 50%.

Summing up, the MIST setup should eliminate the judge's bias from the test results. Its second stage would be unproblematic for judges — it even may be automated. What is more, due to the nature of its questions (addressing commonsense knowledge contributed by non-experts) and the procedure in which they are collected, they should be easy for human participants but difficult for machines (for the same reasons as provided by French for subcognitive questions).

Let us now confront these assumptions, first with the PMI-IR algorithm, and then with the results of the practical evaluation of MIST results.

[22] See French (1990), (2000).
[23] McKinstry (2000).
[24] See McKinstry (1997), (2009).
[25] See McKinstry et al., (2008).

### 3. The PMI-IR algorithm and MIST questions

The PMI-IR algorithm was proposed by Peter D. Turney in his paper "Answering Subcognitive Turing Test Questions: A Reply to French."[26] The PMI-IR stands for Pointwise Mutual Information (PMI) and Information Retrieval (IR). The algorithm measures the semantic similarity between pairs of words or phrases. This involves issuing queries to a search engine and applying statistical analysis to the results. As Turney states: "[t]he power of the algorithm comes from its ability to exploit a huge collection of text."[27] (The technical details of the algorithm are far beyond the reach of this paper, but they are explained in detail in the aforementioned papers.)

The PMI-IR algorithm has been tested against synonym recognition questions retrieved from two standard tests for English learners: TOEFL and ESL.[28] The PMI-IR overall result for TOEFL reached 73.75% (for 80 questions) and for ESL 74% (for 50 questions).

Turney also used the PMI-IR to generate answers to French's subcognitive questions. When the algorithm was applied to the questions retrieved from French's paper[29] it was able to reproduce the expected results. Let us consider one example here, namely the *Flugly* question.

On a scale of 1 (awful) to 10 (excellent), please rate:
- How good is the name Flugly for a glamorous Hollywood actress?
- How good is the name Flugly for an accountant in a W.C. Fields movie?
- How good is the name Flugly for a child's teddy bear?

French expects the following results: "most people would agree that Flugly would be a downright awful name for a sexy actress, a good name for a character in a W.C. Fields movie, and a perfectly appropriate name for a child's teddy bear."[30]

The PMI-IR algorithm assigned the following marks:
- Flugly for a glamorous Hollywood actress = 1;
- Flugly for an accountant in a W.C. Fields movie = 2;
- Flugly for a child's teddy bear = 10.

One may easily notice that they are intuitive and, what is more, in line with French's predictions when it comes to the ranking of the names: actress < accountant < bear (Turney points out that: "[p]erhaps French would give a higher score for Flugly as an accountant, but an informal survey suggests that the above ratings are quite human-like"[31]).

[26] Turney (2001a). The algorithm is also described in (Turney 2001b).
[27] Turney (2001a).
[28] Turney (2001a), (2001b).
[29] French (2000).
[30] French (2000): 336.
[31] Turney (2001a).

Turney was able to repeat such results for other types of questions proposed by French. He concludes the paper in the following way.

> French (1990, 2000) has argued that the Turing Test is too strong, because a machine could be intelligent, yet still fail the test. I agree with this general point, but I disagree with the specific claim that an intelligent but disembodied machine cannot give humanlike answers to subcognitive questions. I show that a simple approach using statistical analysis of a large collection of text can generate seemingly human-like answers to subcognitive questions.[32]

At this point it is also worth mentioning the IBM computer's success in *Jeopardy!* – the American general-knowledge game show. In this show questions could be about anything, and they often rely on complex wordplay. To make things more complicated, the contestant has to supply the correct question to a given clue. A typical example may be: "As an adjective, it means "timely"; in the theatre, it's to supply an actor with a line."[33] The correct response is: "What does "prompt" mean?". In 2011 an IBM computer named Watson defeated two *Jeopardy!* champions Ken Jennings and Brad Rutter.[34] This illustrates the abilities of a modern day AI system in the question processing domain.

The results achieved with the use of the PMI-IR and aforementioned AI success in *Jeopardy!* suggest that there are classes of questions which address commonsense knowledge and which are clearly available for machines. This makes the first assumption of MIST we consider here at least problematic — common sense questions like the one recommended for MIST are easy to answer for humans as well as for modern computers.

## 4. Judge's perspective in MIST

In this section we will take a closer look at the MIST assumption stating that evaluating MIST results would not be problematic for judges. To this end we designed an online study in which one group of participants played the role of a judge in MIST and the other group simply took part in MIST as tested agents. We present the details below.

### 4.1. Methods and Procedure

For our study we used two questionnaires consisting of 50 questions retrieved from the MIST project. As for the selection of questions, we eliminated those which contained vulgarisms and serious grammatical errors, which made them hard to understand (like e.g. "Was thomas nixon born in year?"). Despite this, we have not applied any restrictions on selected questions. Below we present exemplary questions used in the study (associated with the predefined answers retrieved from the MIST project, original spelling is preserved). The complete list of questions may be found in the Appendix of this paper.

[32] Turney (2001a): 419.
[33] See Dormehl (2016): 137.
[34] Dormehl (2016): 138.

| | | |
|---|---|---|
| □ | Is Madonna a woman? | YES |
| □ | Does Santa Claus deliver gifts on Easter? | NO |
| □ | In general, do we need light to see? | YES |
| □ | Is a cell something that can contain either a nucleus or a prisoner? | YES |
| □ | Are most cats furry? | YES |
| □ | Does wood comes from trees? | YES |
| □ | Do all mammals need oxygen to live? | YES |

Both questionnaires were built with the use of the same set of MIST questions. The first questionnaire (hereafter Q1) had the following instruction.

> Your task in this study is to play the role of a judge who is evaluating answers given by a computer program. These answers were given for a simple yes/no questions. Read the question and the answer provided by the program. Afterwards evaluate on a scale 1 (I strongly agree) to 5 (I strongly disagree) the degree in which you agree with the provided answer. If you do not agree with the answer or it is in some sense problematic for you, please give us your comment in the field 'Comment.'

After this instruction, the subject was presented with the list of MIST questions associated with answers, scale for evaluating answers, and a "Comment" text-field.

In the second questionnaire (hereafter Q2) subjects were simply participants of MIST. They were presented with a list of questions and their task was to provide answers ("yes" or "no").

Both Q1 and Q2 ended with questions covering the age, gender and education of subjects. The last question addressed the issue of a rough estimation of computer-use fluency: "Your web-browser started to display many commercials. This makes browsing the internet very hard. What do you do?" Possible answers were: "a) I try to solve it my own" or "b) I try to find someone to solve this problem for me."

Both questionnaires were presented with the use of *Google Forms*.[35] The study was conducted online. Subjects were recruited with the use of social media and internet forums (of a wide topical spectrum, e.g. *Joemonster*, *Wykop*) as we wanted to gather a research group with a variety of subjects. Attention was paid in the recruitment process to ensure that no subject would take fill in both questionnaires. For each address, only one invitation for one questionnaire was sent.

Our main research goal was to check McKinstry's claim that MIST results would be easy for judges. A judge confronted with the MIST result should not have any problems when evaluating answers of subjects. As a measurement of how difficult the evaluation task is, we have chosen Fleiss' Kappa,[36] which is a statistical measure of inter-rater reliability. If this measure is high for the judges group, we can assume that they evaluated MIST answers with a high degree of agreement and consequently that the judgment task was not problematic. Thus, our first research hypothesis is that (H1) for the group of judges (Q1) we would observe a high level of agreement.

[35] http://forms.google.com.
[36] Cf. Carletta (1996).

We apply the same reasoning to the claim that MIST questions are easy to answer for humans. Thus, our (H2) is that also for the group of MIST participants (Q2) we will observe a high level of agreement. For the kappa interpretation, we use values proposed by Viera and Garrett.[37]

### 4.2. Subjects

The research group consisted of 263 subjects. 126 subjects filled out Q1 — i.e., played the role of a judge in MIST. The group consisted with 83% women and 17% men, aged 19–65 (mean=37.93, SD=13.98). The majority of the group had higher education (32%) or were still studying (23%). 82% of subjects pointed the answer (a) to the question about the computer fluency — so they declared that they will try to cope with the browser problem by their own. 137 subjects filled out Q2 — i.e. took part in MIST. The group consisted with 40% women and 60% men, aged 12-67 (mean=28.81, SD=8.68). As for the first group, the majority had higher education (53%). For this group, 93% of subjects declared (a) to be the answer to the question about computer fluency.

### 4.3. Results

For data analysis we used R statistical software.[38] In Table 1 we present Fleiss' Kappa measures for Q1 (MIST judges) and Q2 (MIST participants).

Table 1. The study results — Fleiss' Kappa for Q1 and Q2. Fleiss kappa interpretation by Viera and Garrett[39]

| Questionnaire | N | Fleiss' Kappa | Kappa interpretation |
|---|---|---|---|
| Q1 | 126 | 0.05 | Slight agreement |
| Q2 | 137 | 0.79 | Substantial agreement |

As may be noticed, the first hypothesis was not confirmed. MIST judges reached only *slight agreement* (K=0.05) for their assessments of MIST answers. The result indicates that the task of evaluating a MIST answer is not easy. Several judges confronted with one question and a yes/no answer to this question may disagree on evaluating the answer.

As for the second hypothesis, it is confirmed. The agreement for MIST participants was *substantial*. This suggest that the task of answering MIST questions is rather simple and many subjects confronted with such a question will agree on the answer.

For a better understanding of judges' evaluations for Q1, we also asked our subjects to provide additional explanations. Analysis of these explanations shows that simple answering of questions is not as problematic as evaluating answers. When confronted with such a task, subjects began to analyze the question itself and become more critical. The effect is analogous to the one for the Turing test or the Loebner Contest. There are two distinct tendencies of judges which may be observed in the collected explanations.

[37] Viera and Garrett (2005).
[38] R Core Team (2013).
[39] Viera and Garrett (2005).

The first one refers to a *lack of knowledge*. Certain questions were too specialized or culturally oriented, like "Has SETI discovered extraterrestrial life? [NO];" "Is Idaho in Europe? [NO]." For these question-answer pairs judges often commented "I do not know," "I would first have to check what SETI is." The second tendency is that when confronted with a simple question, judges somehow do not believe that it is that simple. The effect is that even for intuitive questions, like "Does 1 plus 1 equal 3? [NO]" or "Is Napoleon dead? [YES]" subjects try to provide a context, when the answer given is not a proper one. E.g., for the question about Napoleon example comments were the following: "Maybe there is someone else named Napoleon and this person is alive," "Napoleon is alive in history," "My roommate's cat is called Napoleon, and it is all well." There were also questions and answer pairs, which were commented as highly controversial. Many comments of the form "It depends" or "It is controversial" appeared. Examples of such questions are the following: "Do people sometimes lie? [YES];" "Do you need to get a license to have children? [NO];" "Is war better than peace? [NO]."

## 5. Summary

In this paper we focused on one of the most interesting alternatives to the Turing test. McKinstry's Minimum Intelligent Signal Test aims at providing a test better suited for thinking machines. We have evaluated two assumptions made by the MIST author: questions in MIST should be easy for human beings and difficult for machines, and evaluation of MIST results should be non-problematic for judges. When it comes to the first assumption, we have described the PMI-IR algorithm which proved to be able to answer subcognitive questions in a human-like manner. This suggests that a relatively simple statistical approach is effective when it comes to MIST-type questions. On the other hand, the results of our study suggest that these questions are in fact fairly simple for human participants. What is more, the answers gathered for the second questionnaire have a high level of agreement between subjects, which is in line with McKinstry's predictions.

Things are worse when it comes to the judge's role in the MIST. The results of our study indicate that the task of evaluating MIST answers as human-like may be problematic. Subjects who played the role of judges in our study were far from reaching agreement over the answers provided to MIST questions.

Naturally, we are not claiming that the presented results are a conclusive argument against MIST. Our aim was to evaluate the idea and consider its potential weak points. MIST offers a well-defined framework for testing artificial agents. It does not eliminate the judge bias from the test, but certainly the idea of automated evaluation of MIST answers reduces this issue. We also find the idea of using only yes/no question and its justification provided by McKinstry to be a convincing one although this aspect of MIST needs further study. In our opinion, MIST is still one of the best alternatives to TT "on the market" and the most promising one when it comes to potential practical applications. Certainly, the strongest points of the MIST project are the crowdsourcing underlying its questions-responses (*mindpixels*) corpus and the well operationalized idea of the statistical evaluation of a tested agent.

**References**


Ahn L., Blum M., Hopper N.J., Langford J. (2003), "CAPTCHA: Using Hard AI Problems For Security," *Lecture Notes in Computer Science* 2656: 294–311.

Block N. (1995), "The mind as the software of the brain," [in:] *An Invitation to Cognitive Science – Thinking*, E. Smith, D. Osherson, (eds), The MIT Press, London: 377–425.

Carletta J. (1996), "Assessing Agreement on Classification Tasks: The Kappa Statistic," *Computational Linguistics* 22 (2): 249–254.

Dormehl L. (2016), *Thinking Machines. The inside story of Artificial Intelligence and our race to build the future*, WH-Alley, London.

Epstein R., Roberts G., Beber G. (eds) (2009), *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, Springer Publishing Company.

Fetzer J.H. (1995), "Minds and machines: behaviorism, dualism, and beyond," *Stanford Humanities Review* 4 (2): 251–265.

Fetzer J.H. (1997), "Thinking and computing: computers as special kinds of signs," *Minds and Machines* 7 (3): 345–364.

French R. (1990), "Subcogniton and the Limits of the Turing Test," *Mind* 99 (393): 53–65.

French R.M. (2000), "Peeking behind the screen: The unsuspected power of the standard Turing Test," *Journal of Experimental and Theoretical Artificial Intelligence* 12: 331–340.

Garner R. (2009), "The Turing hub as a standard for Turing test interfaces," [in:] *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer,* R. Epstein, G. Roberts, G. Beber (eds), Springer Publishing Company: 319–324.

Harnish R.M. (2002), *Minds, Brains, Computers. An Historical Introduction to the foundations of Cognitive Science*, Blackwell Publishers, Oxford.

Konar A. (2000), *Artificial Intelligence and Soft Computing. Behavioral and Cognitive Modeling of the Human Brain*, CRC Press, Boca Raton–London–N.Y.–Washington.

Loebner H. (2009), "How to hold a Turing test contest," [in:] *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, R. Epstein, G. Roberts, G. Beber (eds), Springer Publishing Company: 173–180.

Łupkowski P. (2010), *Test Turinga. Perspektywa sędziego*, Wydawnictwo Naukowe UAM.

Łupkowski P. (2011), "A Formal Approach to Exploring the Interrogator's Perspective in the Turing Test," *Logic and Logical Philosophy* 20 (1–2): 139–158.

Łupkowski P., Wiśniewski A. (2011), "Turing interrogative games," *Minds and Machines* 21(3): 435–448.

Łupkowski P., Rybacka A. (2016), "Non-cooperative Strategies of Players in the Loebner Contest," *Organon F* 23 (3): 324–365.

Mauldin M.L. (1994), "Chatterbots, Tiny Muds, and the Turing test: entering the Loebner Prize competition," [in:] *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-04)*, Menlo Park (CA): 16–21.

Mauldin M.L. (2009), "Going undercover: Passing as human; artificial interest: A step on the road to AI," [in:] *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer,* R. Epstein, G. Roberts, G. Beber (eds), Springer Publishing Company: 413–430.

McKinstry C. (1997), "Minimum Intelligence Signal Test: an Objective Turing Test," *Canadian Artificial Intelligence* 41: 17–18.

McKinstry C. (2000), "Chris McKinstry Replies: Telescopes, AI And More," URL = https://slashdot.org/story/00/07/04/2114223/chris-mckinstry-replies-telescopes-ai-and--more. [Accessed 12.12.2017].

McKinstry C., Dale R., Spivey M.J. (2008), "Action dynamics reveal parallel competition in decision-making," *Psychological Science* 19 (1): 22–24.

McKinstry C. (2009), "Mind as Space: Toward the Automatic Discovery of a Universal Human Semantic-affective Hyperspace – A Possible Subcognitive Foundation of a Computer Program Able to Pass the Turing Test," [in:] *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer,* R. Epstein, G. Roberts, G. Beber (eds), Springer Publishing Company: 283–300.

Newman A.H., Turing A.M., Jefferson G., Braithwaite R.B. (1952), "Can automatic calculating machines be said to think?", [in:] *The Turing Digital Archive* (www.turingarchive.org), Contents of AMT/B/6.

R Core Team (2013), "R: A language and environment for statistical computing. R Foundation for Statistical Computing," URL = http://www. R-project.org/. [Accessed 20.03.2017].

Saygin A.P., Cicekli I., Akman V. (2001), "Turing test: 50 years later," *Minds and Machines* 10: 463–518.

Shieber S. (ed) (2004), *The Turing Test. Verbal Behavior as the Hallmark of Intelligence*, The MIT Press, Cambridge, Massachusetts, London.

Turing A.M. (1950), "Computing machinery and intelligence," *Mind* LIX (236): 443–455.

Turney D.T. (2001a), "Answering subcognitive Turing test questions: A reply to French," *Journal of Experimental and Theoretical Artificial Intelligence* 13 (4): 409–419.

Turney P.D. (2001b), "Mining the web for synonyms: PMI-IR versus LSA on TOEFL," [in:] *Proceedings of European Conference on Machine Learning*, Springer, Berlin, Heidelberg: 491–502.

Viera A.J., Garrett J.M. (2005), "Understanding Interobserver Agreement: The Kappa Statistic," *Family Medicine* 37 (5): 360–363.

Watt S. (2009), "Can people think? Or machines? A unified protocol for Turing testing," [in:] *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer,* R. Epstein, G. Roberts, G. Beber (eds), Springer Publishing Company: 301–318.

**Appendix: MIST questions used in the study**

| | | |
|---|---|---|
| 1. | Is Madonna a woman? | YES |
| 2. | Is Blackjack a card game? | YES |
| 3. | Does Santa Claus deliver gifts on Easter? | NO |
| 4. | Has SETI discovered extraterrestrial life? | NO |
| 5. | Is wood harder than diamond? | NO |
| 6. | Do some people find genetic engineering to be frightening? | YES |
| 7. | In general, do we need light to see? | YES |
| 8. | Is a cell something that can contain either a nucleus or a prisoner? | YES |
| 9. | Is sun black? | NO |
| 10. | Is air solid? | NO |
| 11. | Is pizza a food for humans? | YES |
| 12. | Is the Milky Way a galaxy? | YES |
| 13. | Does 1 plus 1 equal 3? | NO |
| 14. | Are there over 400 days in a year? | NO |
| 15. | Does one times five equal five hundred? | NO |

| | | |
|---|---|---|
| 16. | Are most cats furry? | YES |
| 17. | Is forward the opposite of backwards? | YES |
| 18. | Is Napoleon dead? | YES |
| 19. | Is it right to take something that is not yours without permission from the owner? | NO |
| 20. | Is Idaho in Europe? | NO |
| 21. | Is this sentence in Spanish? | NO |
| 22. | Is Violet a color? | YES |
| 23. | Is our sun the only star in space? | NO |
| 24. | When you throw a stone in the air, does it keep going up forever? | NO |
| 25. | Does PC stand for "personal computer"? | YES |
| 26. | Do people sometimes lie? | YES |
| 27. | Do humans live on Mars? | NO |
| 28. | Are whales types of fish? | NO |
| 29. | Is Greece a country? | YES |
| 30. | Is the earth as hot as the sun? | NO |
| 31. | Is winter weather warm? | NO |
| 32. | Was Vincent van Gogh a painter? | YES |
| 33. | Is night darker than day? | YES |
| 34. | Is war better than peace? | NO |
| 35. | Is wood the same as metal? | NO |
| 36. | Does a person want to eat when he is hungry? | YES |
| 37. | Is a second shorter than a minute? | YES |
| 38. | Did Germany win WWII? | NO |
| 39. | Is there a maximum number? | NO |
| 40. | Are locks more useful when you have the key? | YES |
| 41. | Is toothpaste a better alternative than sand for brushing teeth? | YES |
| 42. | Do you need to get a license to have children? | NO |
| 43. | Is extraterrestrial life possible? | YES |
| 44. | Does 11 plus 11 equal 22? | YES |
| 45. | Does a week consist of seven days? | YES |
| 46. | Is pregnancy contagious? | NO |
| 47. | Is it safe to drive a car whilst drunk? | NO |
| 48. | Do humans regularly eat other humans? | NO |
| 49. | Does wood comes from trees? | YES |
| 50. | Do all mammals need oxygen to live? | YES |