STANFORD ARTIFICIAL INTELLIGENCE LABORATORY
MEMO AIM-194

STAN-CS-73-347

MULTIDIMENSIONAL ANALYSIS IN EVALUATING A SIMULATION OF PARANOID THOUGHT

BY

KENNETH MARK COLBY

FRANKLIN DENNIS HILF

SUPPORTED BY

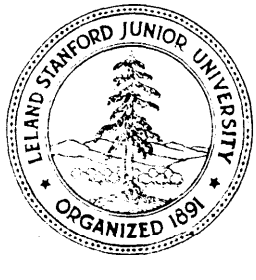
NATIONAL INSTITUTE OF MENTAL HEALTH

MAY 1973

COMPUTER SCIENCE DEPARTMENT

School of Humanities and Sciences

STANFORD UNIVERSITY



STANFORD ARTIFICIAL INTELLIGENCE LABORATORY
MEMO AIM-194

MAY 1973

COMPUTER SCIENCE DEPARTMENT
REPORT NO. CS-347

MULTIDIMENSIONAL ANALYSIS IN EVALUATING A SIMULATION OF PARANOID THOUGHT PROCESSES

by

Kenneth Mark Colby

Franklin Dennis Hilf

ABSTRACT: The limitations of Turing's Test as an evaluation procedure are reviewed. More valuable are tests which ask expert judges to make ratings along multiple dimensions essential to the model. In this way the model's weaknesses become clarified and the model builder learns where the model must be improved.

This research is supported by Grant PHS MH 06645-12 from the National Institute of Mental Health.

Reproduced in the USA. Available from the National Technical Information Service, Springfield, Virginia 22151.

MULTIDIMENSIONAL ANALYSIS IN EVALUATING A SIMULATION OF PARANOID THOUGHT PROCESSES

by

Kenneth Mark Colby
and
Franklin Dennis Hilf

Once a simulation model reaches a stage of intuitive adequacy, a model builder should consider using more stringent evaluation procedures relevant to the model's purposes. For example, if the model is to serve as a as a training device, then a simple evaluation of its pedagogic effectiveness would be sufficient. But when the model is proposed as an explantion of a psychological process, more is demanded of the evaluation procedure.

We shall not describe our model of paranoid processes here. A description can be found in the literature (Colby, Weber, and Hilf, 1971). We shall concentrate on the evaluation problem which asks "how good is the model?" or "how close is the correspondence between the behavior of the model and the phenomenena it is intended to explain?" Turing's Test has often been suggested as an aid in answering this question.

It is very easy to become confused about Turing's Test. In

part this is due to Turing himself who introduced the now-famous imitation game in a paper entitled COMPUTING MACHINERY AND INTELLIGENCE (Turing, 1950). A careful reading of this paper reveals there are actually two imitation games, the second of which is commonly called Turing's Test.

In the first imitation game two groups of judges try to determine which of two interviewees is a woman. Communication between judge and interviewee is by teletype. Each judge is initially informed that one of the interviewees is a woman and one a man uho will pretend to be a woman. After the interview, the judge is asked what we shal l cal l the woman-question i.e. which interviewee was the woman? Turing does not say what else the judge is told but one assumes the judge is NOT told that a computer is involved nor is he asked to determine which interviewee is human and which is the computer. Thus, the first group of judges would interview two interviewees: a woman, and a man pretending to be a woman.

The second group of judges would be given the same initial instructions, but unbeknownst to them, the two interviewees would be a woman and a computer programmed to imitate a woman. Both groups of judges play this game until sufficient statistical data are collected to show how often the right identification is made. The crucial

question then is: do the judges decide wrongly AS OFTEN when the game is played with man and woman as when it is played with a computer substituted for the man. If so, then the program is considered to have succeeded in imitating a woman as well as a man imitating a woman. For emphasis we repeat: in asking the woman-question in this game, judges are not required to identify which interviewee is human and which is machine.

Later on in his paper Turing proposes a variation of the first game. In the second game one interviewee is a man and one is a conipu ter. The judge is asked to determine which is man and which is machine, which we shal I cal I the machine-question. It is this version of the game which is commonly thought of as Turing's test. It has often been suggested as a means of validating computer simulations of psychological processes.

In the course of testing a simulation (PARRY) of paranoid linguistic behavior in a psychiatric interview, we conducted a number of Turing-like indistinguishability tests (Colby, Hilf, Weber and Kraemer, 1972). We say 'Turing-like' because none of them consisted of playing the two games described above. We chose not to play these games for a number of reasons which can be summarized by saying that they do not meet modern criteria for good experimental design. In designing our tests we were primarily interested in learning more

about developing the model. We did not believe the simple machine-question to be a useful one in serving the purpose of progressively increasing the credibility of the model but we investigated a variation of it to satisfy the curiosity of colleagues in artificial intelligence.

In this design eight psychiatrists interviewed by teletype two patients using the technique of machine-mediated interviewing which involves what we term "non-nonverbal" communication since non-verbal cues are made impossible (Hilf, 1972). Each judge interviewed tuo patients one being PARRY and one being a hospitalized paranoid patient. The interviewers were not informed that a simulation was involved nor were they asked to identify which was the machine. Their task was to conduct a diagnostic psychiatric interview and rate each response from the 'patients' along a 0-9 scale of paranoidness, 0 meaning zero and 9 being highest, Transcripts of these interviews, without the ratings of the interviewers, were then utilized for various experiments in which randomly selected expert judges conducted evaluations of the interview transcripts. For example, in one experiment it was found that patients and model were indistinguishable along the dimension of paranoidness.

To ask the machine-question, we sent interview transcripts, one with a patient and one with PARRY, to 100 psychiatrists randomly

selected from the Directory of American Specialists and the Directory of the American Psychiatric Association. Of the 41 replies 21 (51%) made the correct identification while 28 (49%) were wrong. Based on this random sample of 41 psychiatrists, the 95% confidence interval is between 35.9 and 66.5, a range which is close to chance. (Our statistical consultant was Dr. Helena C. Kraemer, Research Associate in Biostatistics, Department of Psychiatry, Stanford University.)

Psychiatrists are considered expert judges of patient interview behavior but they are unfamiliar with computers. Hence we conducted the same test with 100 computer scientists randomly selected from the membership list of the Association for Computing Machinery, ACM. Of the 67 replies 32 (48%) were right and 35 (52%) were wrong. Based on this random sample of 67 computer scientists the 95% confidence ranges from 36 to 60, again close to a chance level.

Thus the answer to this machine-question "can expert judges, psychiatrists aand computer scientists, using teletyped transcripts of psychiatric interviews, distinguish between paranoid patients and a simulation of paranoid processes?" is "No". But what do we learn from this? It is some comfort that the answer was not "yes" and the nuli hypothesis (no differences) failed to be rejected, especially since statistical tests are somewhat biased in favor of rejecting the

null hypothesis (Meehl, 1967). Yet this answer does not tell us what we would most like to know, i.e. how to improve the model. Simulation models do not spring forth in a complete, perfect and final form; they must be gradually developed over time. Perhaps we might obtain a "yes" answer to the machine-question if we allowed a large number of expert judges to conduct the interviews themselves rather than studying transcripts of other interviewers. It would indicate that the model must be improved but unless we systematically investigated how the judges succeeded in making the discrimination we would not know what aspects of the model to work on. The logistics of such a design are immense and obtaining a large N of judges for sound statistical inference would require an effort disproportionate to the information-yield.

A more efficient and informative way to use Turing-like tests is to ask judges to make ordinal ratings along scaled dimensions from teletyped interviews. We shall term this approach asking the dimension-question. One can then compare scaled ratings received by the patients and by the model to precisely determine where and by how much they differ. Model builders strive for a model which shows indistinguishability along some dimensions and distinguishability along others. That is, the model converges on what it is supposed to simulate and diverges from that which it is not.

We mailed paired-interview transcripts to another 400 randomly selected psychiatrists asking them to rate the responses of the two 'patients' along certain dimensions. The judges were divided into groups, each judge being asked to rate responses of each I-O pair in the interviews along four dimensions. The total number of dimensions in this test were twelve - linguistic noncomprehension, thought disorder, organic brain syndrome, bizarreness, anger, fear, ideas of reference, delusions, mistrust, depression, suspiciousness and mania. These are dimensions which psychiatrists commonly use in evaluating patients.

Table 1 shows there were significant differences, with PARRY receiving higher scores along the dimensions of linguistic noncomprehension, thought disorder, bizarreness, anger, mistrust and suspiciousness. On the dimension of delusions the patients were rated significantly higher. There were no significant differences along the dimensions of organic brain syndrome, fear, ideas of reference, depression and mania.

While tests asking the machine-question indicate indistinguishability at the gross level, a study of the finer structure of the model's behavior through ratings along scaled dimensions shows statistically significant differences between patients and model. These differences are of help to the model

builder in suggesting which aspects of the model must be modified and improved in order to be considered an adequate simulation of the class of paranoid patients it is intended to simulate. For example, it is clear that PARRY'S language-comprehension must be improved. Once this has been implemented, a future test will tell us whether improvement has occurred and by how much in comparison to the earlier version. Successive identification of particular areas of failure in the model permits their improvement and the development of more adequate model-versions.

Further evidence that the machine-question is too coarse and insensitive a test comes from the following experiment. In this test we constructed a random version of the paranoid model which utilized PARRY'S output statements but expressed them randomly no matter what the interviewer said. Two psychiatrists conducted interviews with this model, transcripts of which were paired with patient interviews and sent to 200 randomly selected psychiatrists asking both the machine-question and the dimension-question. Of the 63 replies, 34 (49%) were right and 35 (51%) wrong. Based on this random sample of 69 psychiatrists, the 95% confidence interval ranges from 33 to 63, again indicating a chance level. However as shown in Table 2 significant differences appear along the dimensions of linguistic noncomprehension, thought disorder and bizarreness, with RANDOM-PARRY rated higher. On these particular dimensions we can construct a

continuum in which the random version represents one extreme, the actual patients another. our (nonrandom) PARRY lies somewhere between these two extremes, indicating that it performs significantly better than the random version but still requires improvement before being indistinguishable from patients. (See Fig.1). Table 3 presents t values for differences between mean ratings of PARRY and RANDOM-PARRY. (See Table 2 and Fig.1 for the mean ratings).

Thus it can be seen that such a multidimensional analysis provides yardsticks for measuring the adequacy of this or any other dialogue simulation model along the relevant dimensions.

We conclude that when model builders want to conduct tests which indicate in which direction progress lies and to obtain a measure of whether progress is being achieved, the way to use Turing-like tests is to ask expert judges to make ratings along multiple dimensions that are essential to the model. Useful tests do not prove a model, they probe it for its strengths and weaknesses Simply asking the machine-question yields little information relevant to what the model builder most wants to know, namely, along what dimensions must the model be improved.

Table 1. t ratio of correlated means: mean ratings of patient I-O pairs vs mean ratings of PARRY I-O pairs.

Dimension	n of Judges	Mean Patient Ratings	Mean PARRY Ratings	Mean Deviation	Standard Error of Difference	t
Linguistic Non-Comprehension	43	0.73	2.22	-1.50	0.28	-5.28**
Thought Disorder	43	2.29	3.78	-1.49	0.41	-3.60**
Organic Brain Syndrome	43	0.84	1.11	-0.27	0.29	-0.93
Bizarreness	42	2.34	3.45	-1.19	0.36	-3.28*
Anger	37	2.03	2.96	-0.92	0.21	-4.30**
Fear	38	2.73	2.67	0.06	0.22	0.26
Ideas of Reference	36	2.33	1.78	0.55	0.32	1.71
Delusions	37	3.06	1.51	1.55	0.33	4.70**
Mistrust	41	2.35	4.42	-2.13	0.35	-6.14**
Depression	39	1.92	1.46	0.25	0.21	1.21
Suspiciousness	40	2.87	4.33	-1.43	0.36	-3.98**
Mania	40	1.00	1.23	-0.09	0.29	-0.32

*Level of significance better than .01.

**Level of significance better than .001.

Table 2. t ratio of correlated means: mean ratings of patient I-O pairsvs mean ratings of RANDOM-PARRY I-O pairs.

Dimension	n of Judges	Mean Patient Ratings	Mean RANDOM PARRY Ratings	Mean Deviation	Standard Error of Difference	t
Linguistic Non-Comprehension	25	0.51	2.83	-2.30	0.51	-4.51**
Thought Disorder	26	2.99	5.94	-2.96	0.36	-8.18**
Organic Brain Syndrome	25	0.87	1.19	-0.32	0.36	-0.89
Bizarreness	26	2.38	4.89	-2.50	0.41	-6.05**

*Level of significance better than .01.

**Level of significance better than .001.

Table 3. t values for difference between independent means: mean ratings of PARRY vs RANDOM-PARRY. A minus value of t indicates that RANDOM-PARRY is higher.

Dimension	Degrees of Freedom	t	Level of Significance of Difference
Linguistic Non-Comprehension	66	-1.39	not significant
Thought Disorder	67	-3.87	.001
Organic Brain Syndrome	66	-0.19	not significant
Bizarreness	67	-2.76	.01

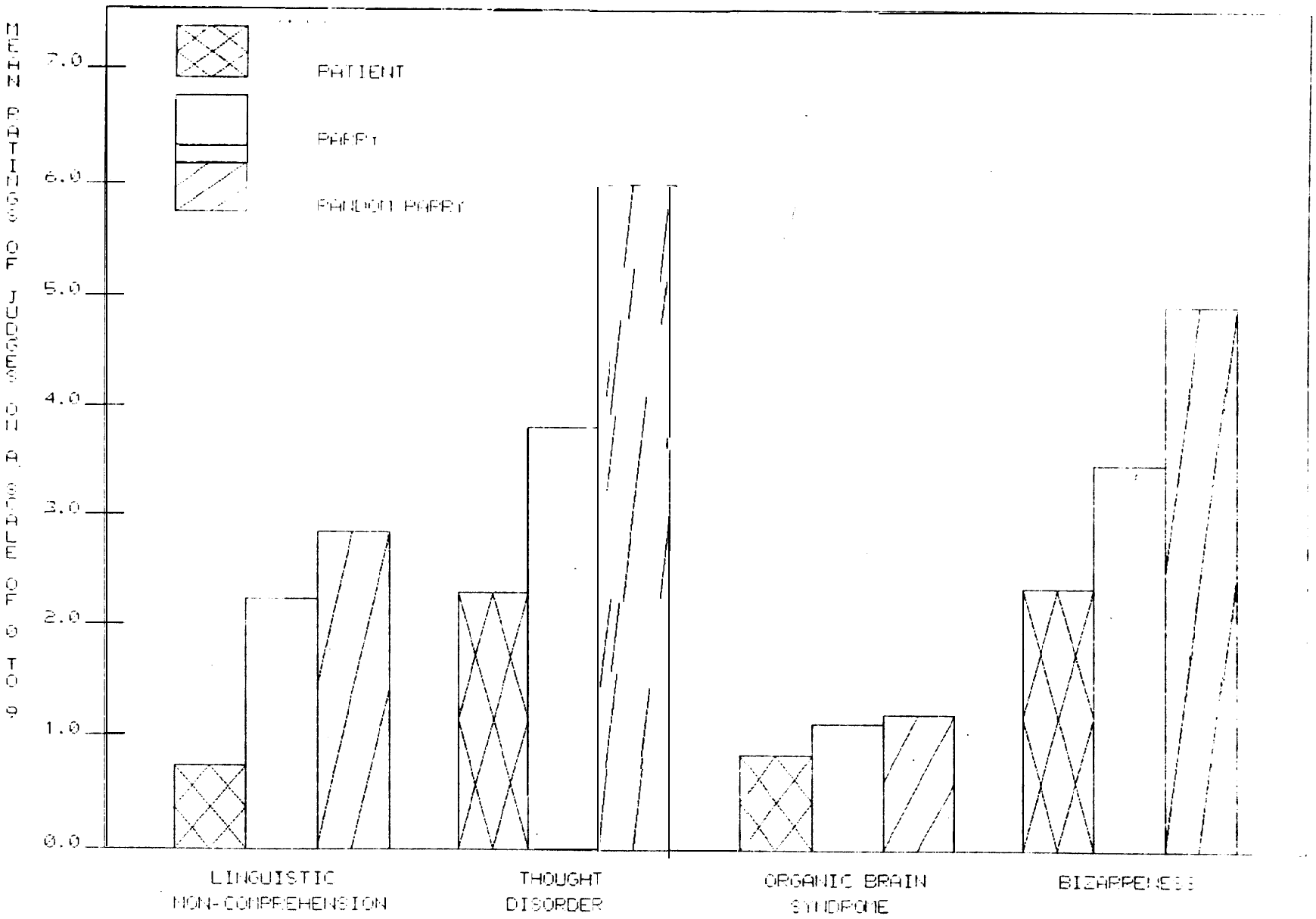


Fig. 1: IN ALL 4 DIMENSIONS THE PARRY RATINGS ARE CLOSER TO THE PATIENT RATINGS THAN ARE THE RANDOM PARRY RATINGS.

REFERENCES

[1] Colby, K.M., Weber, S. and Hilf, F.D., 1971. Artificial paranoia. ARTIFICIAL INTELLIGENCE,& I-25.

[2] Colby, K.M., Hilf, F.D., Weber, S. and Kraemer, H.C., 1972. Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. ARTIFICIAL INTELLIGENCE, 3, 199-221.

[3] Hilf, F.D., 1972. Non-nonverbal communication and psychiatric research. ARCHIVES OF GENERAL PSYCHIATRY, 27, 631-635.

[4] Meehl, P.E., 1967. Theory testing in psychology and physics: a methodological paradox. PHILOSOPHY OF SCIENCE, 34, 103-115.

[5] Turing, A., 1950. Computing machinery and intelligence. Reprinted in: COMPUTERS AND THOUGHT (Feigenbaum, E.A. and Feldman, J., eds.). McGraw-Hill, New York, 1963, pp. 11-35.