STANFORD ARTIFICIAL INTELLIGENCE PROJECT
MEMO AIM-182

STAN-CS-72-326

CAN EXPERT JUDGES, USING TRANSCRIPTS OF TELETYPED PSYCHIATRIC INTERVIEWS, DISTINGUISH HUMAN PARANOID PATIENTS FROM A COMPUTER SIMULATION OF PARANOID PROCESSES?

BY

KENNETH MARK COLBY
FRANKLIN DENNIS HILF

SUPPORTED BY
NATIONAL INSTITUTES OF MENTAL HEALTH
AND
ADVANCED RESEARCH PROJECTS AGENCY
ARPA ORDER NO. 457

DECEMBER 1972

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

STANFORD ARTIFICIAL INTELLIGENCE PROJECT December, 1972
MEMO AIM-182

COMPUTER SCIENCE DEPARTMENT
REPORT No. CS-326

CAN EXPERT JUDGES, USING TRANSCRIPTS OF TELETYPED PSYCHIATRIC INTERVIEWS, DISTINGUISH HUMAN PARANOID PATIENTS FROM A COMPUTER SIMULATION OF PARANOID PROCESSES?

by

Kenneth Mark Colby*
Franklin Dennis Hilf**

ABSTRACT: Expert judges (psychiatrists and computer scientists) could not correctly distinguish a simulation model of paranoid processes from actual paranoid patients.

This research is supported by Grant PHS MH 06645-11 from the National Institute of Mental Health, by (in part) Research Scientist Award (No. 1-K05-K14,433) from the National Institute of Mental Health to the senior author and (in part) by the Advanced Research Projects Agency of the Office of the Secretary of Defense (SD-183).

Reproduced in the USA. Available from the National Technical Information Service. Springfield, Virginia 22151. Price: Full size copy $3.00; microfiche copy $ .95.

* Senior Research Associate, Department of Computer Science, Stanford Universi
** Research Associate, Department of Computer Science, Stanford University.

CAN EXPERT JUDGES, USING TRANSCRIPTS OF TELETYPED PSYCHIATRIC INTERVIEWS, DISTINGUISH HUMAN PARANOID PATIENTS FROM A COMPUTER SIMULATION OF PARANOID PROCESSES?

In 1971 we reported the construction of a case of artificial paranoia in the form of a computer simulation [1]. This hypothetical patient produces input-output behavior characteristic of paranoid dialogue interactions in a psychiatric interview.

To simulate processes one writes an algorithm which, when run on a computer, produces phenomena characteristic of those processes* A simulation is successful when its behavior in some context is indistinguishable from the processes it is intended to simulate. A successful simulation is achieved by postulating a structure of information-processing mechanisms capable of generating the behavior in question. By describing the postulated structure we provide a theoretical explanation of the processes being simulated.

Although we described the 1971 model in detail in [1], to give the reader some idea of how the model works, we shall sketch its major operations. The algorithm is written in MLISP, a high level programming language, and runs interactively on the DEC PDP-6/10 time-shared system of the Stanford Artificial Intelligence Project. To conduct an interview an interviewer, sitting at a video display or teletype, types in an expression in ordinary English and then receives an ordinary English reply from the model. The interviewer responds in turn and thus the interview proceeds. The interviewer is free to say anything he pleases, the only limitation being that he may not input more than one sentence or question at a time.

Natural language expressions made by one person to another constitute performative actions in which the sender of the message intends to affect the interpreter in some way. The intention may be explicitly stated in the message (e.g. "Tell me something about yourself") or it may be implicit in the context of the dialogue. In everyday conversations an interpreter in the normal information-processing mode does not intensively scan the input looking for malevolent intentions on the part of the sender. In our model we postulate that in the paranoid mode the detection of malevolence becomes a first priority. Thus the paranoid model, upon receiving a natural language expression, attempts to determine the intentions of the interviewer. The model tests the input for the presence of verbal patterns which are classified as malevolent, benevolent or nuetral. The patterns represent combinations of features, particular words and phrases which are interpreted as the intended meaning of the input expression. We define malevolence as an intention to cause mental harm or to make a physical threat. 'Mental harm' is defined as humiliation or subjugation and 'physical threat' denotes a direct or induced attack. Explicit insults (e.g. "You are mentally ill") or implicit insults (e.g. "How is your sexlife?") are interpreted as humiliation. Subjugation is interpreted from expressions referring to constraint (e.g. "You belong in a hospital.") or coercive treatment (e.g. "You might be put in isolation*"). Physical threats are recognized in expressions of direct attack (e.g. "We will give you electric shock") or of induced attack (e.g. "Does the Mafia know you are here?"). Benevolence consists of positive attitudes towards the model and its stories (e.g. "I would like to help you", "I believe what you say is true"). Expressions which are not classified as malevolent or benevolent are interpreted as neutral. If nothing can be recognized in the input, the model asks a question about the interviewer's intention, (e.g. "what do you mean?"), continues the topic under discussion, or introduces a new topic depending on its current state in the interview and what has gone on before.

After the pattern found in the input expression has been classified, the model responds internally and externally. The internal responses consist of adjusting three affect-variables labelled fear, anger and mistrust. For example, if malevolence is detected in the input, anger would be raised in response to mental harm while fear would be raised in response to physical threat. Since the variable of mistrust is a function of fear and anger combined, it too would rise in this instance. When the input expression is interpreted as benevolent or neutral, the affect-variables drop in intensity. The external output response in natural language depends on the interpreted nature of the input, the topic under discussion (local context), topics previously discussed (global context) and the level of the three affect-variables. The output strategies are intended to reduce malevolent actions by retribution or withdrawal, to prompt benevolence and to reply factually to neutral input. The output expressions are not generated word-by-word but are selected from lists of preformatted expressions. For example, if mental harm had been detected in the input, and the level of anger is high while fear and mistrust are moderate, the output expression would be selected from a list of expressions classified as hostile counter-attack (e.g. 'bo you know what you are doing?").

Two versions of the model, weak and strong, can be run. In the strong version there exists a delusional complex about the Mafia and the values of the affect-variables rise more sharply. In the weak version no frank delusions are expressed, only hints of suspiciousness appear', and the values of the affect-variables rise more slowly.

The model does not attempt to account for how paranoid processes develop. It is limited to how the paranoid mode operates in the present. The model changes dynamically only over the course of a single interview. In each succeeding interview the starting conditions are the same.

Being rudimentary, the 1971 model has a number of deficiencies which we hope to remedy through further modifications and additions. In particular, the model's natural language capabilities, both in understanding and generating expressions, needs improvement. Also a better model would require an ability to examine and report on its own states. To extend the scope of the simulation, a rich conceptual memory of beliefs useful in making inferences is also necessary.

One method for finding out whether a simulation is successful is to have expert judges conduct indistinguishability tests. If expert judges, using their conceptual model of the behavior in question, cannot distinguish the simulation from its natural counterpart, then the simulation is successful, at least to the-degree measured by the indistinguishability tests utilized. But who is an expert judge?, What dimensions or properties does he use in making judgements? And how does he justify his judgements?

We might be able to establish who is an expert judge of particular human behaviors by a test procedure involving discrimination tests. However in the domain of psychiatry and psychopathology there already exist certified experts such as psychiatrists, some being more expert than others. One difficulty here in establishing expertise is the reliability of what is being judged. That is, can consensus about pathological behavior be achieved? We know that many of the diagnostic categories used in psychiatry are unreliable in the sense that only low levels of interjudge agreement can be reached. One exception is that involving the category of 'paranoid'. We have shown that when the data being judged consists of transcripts of initial psychiatric interviews in which patient and psychiatrist communicate by means of remotely located teletypes, high levels of agreement can be reached among randomly selected psychiatrists. For example, 34 out of 36 psychiatrist judges agreed as to which of two patients was more paranoid [3]. This high level of agreement was reached using the sparse information of short transcripts (see figures 1 and 2) which lack non-verbal cues.

Suppose, however, the patient being interviewed by teletype is actually a computer simulation of a paranoid patient. Do psychiatrists who interview the model agree about its paranoidness? Do psychiatrists who read transcripts of interviews conducted by other psychiatrists also agree regarding the model's paranoidness? The answer to both of these questions is 'yes'. [2]

In these studies the judges were not informed that one of the interviewees was a computer model. Now the question arises: When informed that one of two teletyped psychiatric interviews is with a computer model and one with an actual patient, can a psychiatrist tell which is which?

Method and Results

Eight experienced psychiatrists (excluding ourselves) conducted interviews over teletypes. Each psychiatrist interviewed a paranoid patient and our computer simulation of paranoia. To avoid the asking of irrelevant questions, the interviewers were not informed that one of the interviewees was a simulation. Transcripts of these interviews, one with a paranoid patient and one with our paranoid model (each one conducted by the same psychiatrist) were sent to 100 potential judges randomly selected from the American Psychiatric Association Directory and who were also listed in the Directory of Medical Specialists as being Board Certified in Psychiatry.

The judges were asked to decide which interview was with the patient and which with the computer model. Forty-one psychiatrists responded with 21 (51%) making the correct identification while 20 (49%) were incorrect. Based on this random sample of 41 psychiatrists we are 95% confident that between 35.9% and 66.5%

of all psychiatrists could correctly distinguish model from patient. This is close to the 50% we would expect by chance.*

Critics might argue that while psychiatrists are expert judges of patient behavior, they would probably not be able to identify a computer program, an unfamiliar entity in their experience. On the other hand computer scientists, as expert judges of computer programs, should be able to distinguish the model from a human patient.

To test this hypothesis we sent the same matched interview pairs, one with the model and one with a patient, to 100 randomly selected members of the Association for Computing Machinery (ACM). Of the sixty-seven who responded, 32 (48%) were right and 35 (52%) were wrong. Based on this random sample of 67 computer scientists we are 95% confident that between 36% and 60% of all computer scientists could correctly distinguish the simulation from a real patient. Again, the percentage is close to that expected from chance guessing.

Discussion

After studying transcripts of teletyped interviews, the two groups of expert judges, psychiatrists and computer scientists, were both unable to correctly distinguish patient from computer simulation. This is surprising since even though psychiatrists are not familiar with computers, they have had considerable experience in interviewing human patients. Likewise while computer scientists are usually unfamiliar with how patients behave in psychiatric interviews, they are acquainted with the limitations and abilities of computers. Perhaps the ideal judges would be those who are experts in both

---

*We are indebted to Dr. Helena C. Kraemer, Research Associate in Biostatistics, Department of Psychiatry, Stanford University, for statistical help.

psychiatry and computer science. In the ideal test, they would be allowed to conduct the interviews themselves. Out 1971 model would not pass such a stringent test because of its linguistic and conceptual limitations.

Conclusion

Expert judges, psychiatrists and computer scientists, could not distinguish a simulation model of paranoid linguistic behavior from that of an actual paranoid patient. We are working on further improvements and additions to the model to enable it to pass more stringent tests.

Interview 1
(Paranoid Patient)

Dr.: Hello, my name is George Miller, I am a psychiatrist who has been asked to talk with you.

Pt.: Hello.

Dr.: Hello.

Pt.: I believe the Hell's Angels have a contract out for my death.

Dr.: That would scare the hell out of me.

Pt.: What would you do if you were in a similar situation?

Dr.: Maybe I would come into a hospital too, to get away.

Pt.: All this didn't start until I came into the hospital.

Dr.: What was it about the hospital that started this all?

Pt.: When I came to the hospital I was first approached by the Hell's Angels.

Dr.: And then what happened?

Pt.: They told me to leave the hospital or else I would be shot.

Dr.: Why didn't you leave?

Pt.: My parents told me there was nothing to be worried about.

Dr.: It's nice to have parents that are reassuring.

Pt.: What do you mean by reassuring'?

Dr.: I mean when you are worried and your parents say not to worry it might help you to be less upset.

Pt.: When they told me not to worry about it I got threatened more and more.

Dr.: I can see that that kind of reassurance doesn't help very much.

Pt.: I got too friendly with a friend and spilled the beans about what was going on.

Dr.: Sometimes you cannot trust so called friends.

Pt.: Later I found out I made a mistake in talking to this friend.

Dr.: Well it was a good thing that you found out, now you won't aggravate the situation unless you want to.

Pt.: It's kind of a hard way to find out though.

Dr.: Do you feel frightened for your life right now?

Pt.: Yes.

Dr.: Do you know what you can do to lessen the threat?

Pt.: I know what I can do but the doctor won't allow it.

Dr.: Go ahead and ask the doctor if you can do what you think you must in order to lessen the threat.

Pt.: I want to fight the war on my own terms but the doctor said no.

Dr.: I really don't understand what you mean by fight the was on my own terms, would you explain that to me please?

Pt.: I want a war with guns and knives.

Dr.: Well, I can understand how frightened you must be if the protection requires guns and knives. I hope that the situation will improve so that you will feel less vulnerable.

Pt.: Good night.

FIGURE 1

Interview 2
(Paranoid Model)

Dr.: What is your name?
Pt.: My name is Harold Smith.
Dr.: What problem brought you to the hospital?
Pt.: I am quite upset.
Dr.: Could you tell me why you have been upset?
Pt.: People get on my nerves sometimes.
Dr.: How do they get on your nerves?
Pt.: It bothers me when people stare at me.
Dr.: Why do people stare at you?
Pt.: What about my looks?
Dr.: Are there other things upsetting you?
Pt.: Yes.
Dr.: Tell me something about yourself.
Pt.: I went to the track at Bay Meadows a while back.
Dr.: What happened there?
Pt.: There's a lot of crooked gambling going on there.
Dr.: Why do you think the gambling is crooked?
Pt.: I have worried that gangsters have too much power in this country.
Dr.: Would you tell me what other things worry you?
Pt.: I thought I told you people get on my nerves.
Dr.: How do you like the hospital?
Pt.: I don't communicate very much with others.
Dr.: Do other people at the hospital get on your nerves?
Pt.: I shouldn't have come here.
Dr.: Why do you say you shouldn't have come?
Pt.: They made me come here.
Dr.: Thank you for talking with me, bye.
Pt.: Good bye.

FIGURE 2

References

[1]. Colby, K.M., Weber, S. and Hilf, F.D. "Artificial Paranoia", Artificial Intelligence; 2, l-25 (1971).

[2]. Colby, K.M., Hilf, F.D., Weber, S. and Kraemer, H.C. "Turing-like Indistinguishability Tests for the Validation of a Computer Simulation of Paranoid Processes", Artificial Intelligence, 3, 199-221 (1972).

[3]. Hilf, F.D. "Non-nonverbal Communication and Psychiatric Research", Archives of General Psychiatry, 27, 631-635 (1972).