STANFORD ARTIFIC IAL INTELLIGENCE PROJECT
MEMO A IM-156

COMPUTER SC IENCE DEPARTMENT
REPORT NO. CS-246

A RESEMBLANCE TEST FOR THE VALIDATION OF A COMPUTER
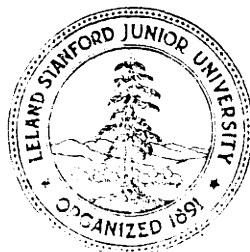SIMULATION OF PARANOID PROCESSES

BY

KENNETH MARK COLBY

FRANKLIN DENN IS H I LF

SYLVIA WEBER

HELENA C. KRAEMER

NOVEMBER 1971

COMPUTER SCIENCE DEPARTMENT
STANFORD UNIVERSITY

A RESEMBLANCE TEST FOR THE VALIDATION OF A

COMPUTER SIMULATION OF PARANOID PROCESSES

by

Kenneth Mark Colby*

Franklin Dennis Hilf**

Sylvia Weber†

Helena C. Kraemer††

ABSTRACT: A computer simulation of paranoid processes in the form
            of a dialogue algorithm was subjected to a validation
            study using an experimental resemblance test in which
            judges rated degrees of paranoia present in initial
            psychiatric interviews of both paranoid patients and
            of versions of the paranoid model. The statistical
            results indicate a satisfactory degree of resemblance
            between the two groups of interviews. It is concluded
            that the model provides a successful simulation of
            naturally occuring paranoid processes.

* Senior Research Associate, Department of Computer Science, Stanford University.
** Research Associate, Department of Computer Science, Stanford University.
 † Graduate Student, Department of Ccnnputer Science, Stanford University.
†† Research Associate in Biostatistics, Department of Psychiatry, Stanford
   University.

A RESEMBLANCE TEST FOR THE VALIDATION OF A
COMPUTER SIMULATION OF PARANOID PROCESSES

Introduction

Those of us who work in the area of **computer simulation of human**
mental functions have been concerned for some time about how faithful
our representations are of the represented processes. Sympathetic
critics have complained that, while simulation models may be imaginative
and interesting, insufficient attention has been paid to the problem
of how well or to what degree a given model corresponds to the **modelled**
process.

While reasonably satisfying replies to this criticism have not yet
been offered, some defense has been attempted in the form of postponing
tactics. One line of defense calls on the nature of historical phases
in scientific inquiry. It is routine in analyzing scientific activity
to make a distinction between an initial phase of invention or discovery
and a second phase called 'methodology' in which justification of the
first phase is attempted through validation procedures. In the initial
phase one must explore, invent, construct models, theorize, etc. without
becoming paralyzed by worry over the outcomes of second phase. As
Medawar [7] has succinctly put it:

> "Too much can be made of matters of validation. Scientific
>
> research is not a clamor of affirmation and denial. Theories
>
> and hypotheses are modified more often than they are discredited.
>
> A realistic methodology must be one that allows for repair as
>
> readily as for refutation."

Model-builders first try to construct a model which has at least

intuitive adequacy by rough criteria of face-validity and a **complex**

simulation may initially require thousands of hours of coding and

debugging before this adequacy is achieved. The day then comes when

the methodological phase is reached in which attempts to validate the

model by means of critical tests **must be** considered.

A second line of defense for model-builders has been that there

are as yet no satisfactory procedures for validating simulation models.

That is, the first phase of inventive work in validation procedures for

these sorts of models has not progressed even to intuitive adequacy. A

number of proposals for validation have been made (they are reviewed in

Abelson [1]) but they have thus far not gained widespread trials. One

exception is **Newell's** extensive experience with matching responses of

thinking-aloud protocols with program traces. There are difficulties

with this sort of canparison, as Newell [8] points out:

> "Although a human can assess each instance qualitatively, there
>
> are no available techniques for quantifying the canparison, or
>
> **summarizing** the results of a large set of **camparisons.**"

These attempts to reply to our critics have been accepted or re-

jected depending on degrees of goodwill. A serious model-builder

realizes that his syntheses must eventually be empirically tested and,

if there are no testing procedures available, he should consider **de-**

.veloping some. After all, he is in the best position to know what the

requirements of appropriate evaluation should be.

Validation of Computer Simulations

The term 'validation' is used in a number of ways in scientific and

philosophical writings. To some scientists validity is a truth-status

attribute of theories; to all logicians validity is an attribute of
deductive argument.

'Validate' derives from the Latin **validus** = to be strong. To
validate X would be to add strength, weight, force or convincingness
to the acceptability of X. But acceptability to whom and as what?
One validates X both for oneself and for that small community of an
expert forum capable of judging the acceptability of X. Since X in
our case is a simulation model, its **acceptability** as a simulation must
first be based on its success in achieving the desired end of producing
resemblance at some input-output level. An acceptable simulation is
one which succeeds, according to some relevant test, at input-output
imitation. To evaluate the success of a given simulation is to subject
it to a test procedure which it can pass or (more importantly) fail.
The evaluation of the acceptability of a simulation as a successful
imitation is a different problem from the evaluation of a simulation
as an acceptable model-explanation.

To determine the degree of resemblance produced by a simulation
one utilizes experimental tests. If a simulation is not judged to be different
from its natural counterpart along certain dimensions, then the simulation
can be considered successful. It is presupposed in this argument that
there are stipulated dimensions of the resemblance and that there exist
-relevant test operations in making judgements of similarity and difference.

In 1637 Descartes [4] proposed two tests for distinguishing men from
machines designed to resemble them:

"If there were machines which bore a resemblance to our body
and imitated our actions as far as it was morally possible to
do so, we should always have two very certain tests by which

to recognize that, for all that, they were not real men.

The first is, that they could never use speech or other signs as we do when placing our thoughts on record for the benefit of others. For we can easily understand a machine's being constituted so that it can utter words, and even emit some responses to action on it of a corporeal kind, which brings about a change in its organs; for instance, if it is touched in a particular part it may ask what we wish to say to it; if in another part it may exclaim that it is being hurt, and so on. But it never happens that it arranges its speech in various ways, in order to reply appropriately to everything that may be said in its presence, as even the lowest type of man can do. And the second difference is, that although machines can perform certain things as well as or perhaps better than any of us can do, they infallibly fall short in others, by the which means we may discover that they did not act from knowledge, but only from the disposition of their organs. For while reasons is a universal instrument which can serve for all contingencies, these organs have need of some special adaption for every particular action. From this it follows that it is morally impossible that there should be sufficient diversity in any machine to allow it to act in all the events of life in the same way as our reasons causes us to act."

Another type of resemblance test was suggested in 1950 by Turing [9] who termed it 'The Imitation Game'. Since there is so much confusion about Turing's Test (some say it proves machines can think and some

claim it is a good test of a simulation), the relevant portion of
Turing's description will be quoted at length:

'I propose to consider the question, 'Can machines think?'
This should begin with definitions of the meaning of the
terms 'machine' and 'think'. The definitions might be framed
so as to reflect so far as possible the normal use of the
words, but this attitude is dangerous. If the meaning of
the words 'machine' and 'think' are to be found by examining
how they are commonly used it is difficult to escape the
conclusion that the meaning and the answer to the question,
'Can machines think?' is to be sought in a statistical
survey such as a Gallup poll. But this is absurd. Instead
of attempting such a definition I shall replace the question
by another, which is closely related to it and is expressed
in relatively unambiguous words.

The new form of the problem can be described in terms
of a game which we call the 'imitation game'. It is played
with three people, a man (A), a woman (B), and an interrogator
(C) who may be of either sex. The interrogator stays in a
room apart from the other two. The object of the game for the
interrogator is to determine which of the other two is the
man and which is the woman. He knows them by labels X and Y,
and at the end of the game he says either 'X is A and Y is B'
or 'Xis B and Y is A.' The interrogator is allowed to put
questions to A and B thus:

c: Will X please tell me the length of his or her hair?
Now suppose X is actually A, then A must answer. It is

A's object in the game to try and cause C to make the wrong

identification. His answer might therefore be:

'My hair is shingled, and the longest strands are about

nine inches long.'

In order that tones of voice may not help the interrogator

the answers should be written, or better still, typewritten.

The ideal arrangement is to have a teleprinter communicating

between the two rooms. Alternatively the question and answers

can be repeated by an intermediary. The object of the game for

the third player (B) is to help the interrogator. The best

strategy for her is probably to give truthful answers. She can

add such things as 'I am the woman, don't listen to him!' to

her answers, but it will avail nothing as the man can make

similar remarks.

We now ask the question, 'What will happen when a machine

takes the part of A in this game?' Will the interrogator decide

wrongly as often when the game is played like this as he does

when the game is played between a man and woman? These questions

replace our original, 'Can machines think?' "

As an experimental design for a validation procedure there are a

number of weaknesses in the game Turing proposed. The dimension of

'womanliness' is too vague a conceptual dimension to make a judgement

about using purely linguistic information. There are no known criteria

for identifying women over teletypes. An ability to deceive on the

part of a man is required and the ordinary man may have no skill at

this. (Why not use professional female impersonators? But are they

really men?) Finally, since the variable is dichotomous, if a computer

fails to imitate a man imitating a woman, then is it a successful

imitation of a man? From these considerations, we conclude that the simple Imitation Game is a weak test.

It is not stated in Turing's description exactly what the interrogator is told before playing the game. What a judge is told and what he believes is happening are extremely important in resemblance tests. From our experience we have learned that it is unwise to inform a judge at the start that a-computer simulation is involved because he then tends to ask questions designed to detect which of two respondents is a program rather than asking questions relevant to the dimension to be identified. Of course there is no way to prevent a judge from having the fantasy he is communicating with a program.

Some **people** seem confident that they can distinguish a **program**-respondent from a human-respondent by conversational means. But it is not so easy when the human-respondent does not adopt tacit conversational rules a judge expects him to abide by. Some of these rules as standards appropriate to the context involve answering honestly and candidly to the best of one's knowledge, not to be sarcastic and not replying in a joking mode. We have found in informal experiments that if a **human**-respondent does not follow standards of the interviewer's expectations, jokes around, or plays other games, ordinary judges cannot distinguish him from a **computer** program with limited natural language understanding.

Improving on the 'simple Turing Test' of the Imitation Game, Abelson [1] proposed an 'Extended Turing Test'.

"As before, there is a canputer program intended as an imitation of a subject carrying out a set of tasks. But there is also another target person whom we may designate the <u>foil</u>. <u>The foil differs</u> <u>from the subject with respect to some simple dimension</u>, e.g., sex,

age, skill, or etc. In a series of baseline runs of the Game, the subject works in one room and the foil in the second. The judge, using typewritten output, must guess the correct identities of subject and foil; for example, which is the man and which the woman. Over a series of runs, the judge will guess correctly some percentage of the time. For illustrative purposes, suppose that this base percentage is 70 percent. At some point in the procedure, a computer program is substituted for the subject while the foil continues as before. The judge must again guess the correct identities of, e.g., the 'man' and the 'woman'. Turing does not make clear whether the judge is ever told anything at all about the entrance of a computer into the game. The best procedure is undoubtedly not to inform him, and to interlace subject vs. foil runs with computer vs. foil runs. As far as the judge knows, he must on every run look for cues relevant to the announced dimension of difference between subject and foil. The crucial datum is the percentage of correct identifications of the foil when pitted against the computer-simulated 'subject'. Denote this percentage the <u>test percentage.</u> The simulation is judged successful if both base and test percentages reliably exceed 50 percent and the test percentage is not statistically different from the base percentage. Such a success (or failure) would, however, only partially validate (or invlaidate) the simulation. This validation test is relative to the dimension of difference between subject and foil. For illustrative purposes, suppose that in a problem-solving task situation with a man as subject

and a woman as foil, the base percentage were 70 percent and

the test percentage 90 percent. That is, the computer's

task protocols are more easily distinguished from the woman's

protocols than the man's protocols are from the woman's, In

view of the fact that the judge-is told nothing about computers,

only to distinguish the man's work from the woman's, such a

result indicates that the computer behaves in a manner which

is 'too male'. This would come about if the computer protocols

contained an overabundance of some stereo-typically male attribute

such as an analytic rather than intuitive approach to the

problem task. On the other hand, if the test percentage were

55 percent-; then this indicates that the computer program is

'not male enough'. In either case, a particular kind of change

in the simulation is indicated. Finally, supposing the test

percentage to have been 67 percent, or 72 percent, or anything

within a statistically acceptable range around 70 percent, then

the simulation is judged acceptable with respect to its maleness.

The investigator might then wish to proceed to another validation

test using a different foil dimension, perhaps intelligence, or

experience with problem-solving, etc. Our version of an Extended

Turing Test definitely is meant to require the use of several

foil dimensions (and also foils at different positions along

continuous dimensions such as intelligence) before the simulation

can be considered validated."

Modifying Abelson's proposal, we devised an experimental design suitable

for testing our simulation of paranoid processes which consists of an algorithm capable of participating in natural language dialogue[1] . Our purpose was <u>not</u> to play Turing's Game of identifying which respondent is a computer. As in Abelson's 'Extended Test', the task of **the** expert judges (psychiatrists) in our test was to rate degrees along a dimension, in this case the specific process termed 'paranoid'. The conceptual dimension 'paranoid' has the advantage of being one of the few reliable categories (85-95 percent agreement) in the current psychiatric **clas**-sification scheme, not only for ratings of presence or absence but also for ratings of severity, [2]. In our experiment the judges interviewed paranoid patients by means of remote teletype messages. Versions of a paranoid simulation model were included as 'patients' to be interviewed. These versions represented two separate positions, weak and strong, along the paranoid dimension. Weak and strong versions of the model were utilized to see if our control over the degree of paranoia exhibited by the model corresponded with judgements of severity made by psychiatrists. The workings of the model will not be described here. They have been presented in detail elsewhere [3].

In constructing and improving a model one repeatedly uses 'face validity' procedures in striving for intuitive adequacy. That is, the model builders check the behavior of the model against their own conceptual representation of the **modelled** processes for faithfulness of the model's input-output resemblance. In creating simulations it is assumed that the model builders possess a fair idea of what the **modelled** process looks like at the level of input-output behavior and can make a judgement that the simulation's I-O behavior corresponds to some degree to the I-O behavior

of the process being simulated. Another aid in this face-validity pro-
cedure is to have experts with extensive **knowledge** of the **modelled** pro-
cess judge how well the model performs along the specified dimension of
the simulation. In the case of our model, would practicing clinical
psychiatrists, who have credentials as experts, judge its communicative
behavior to be paranoid and would judgements of severity correlate with
the weak and strong versions of the model?

In the early states of constructing the dialogue algorithm dozens of
psychiatrists and clinical psychologists interviewed versions of the
model and we continue to demonstrate them informally to visiting clinicians.
From this experience we learned where and how to improve the model's behavior.
A successful simulation should converge on what it is imitating and diverge
from what it is not trying to imitate. For example, if a model is not in-
tended to simulate chronic brain damage, its I-O behavior should not be
judged as such. Over a period of about a year all but three of this
group of clinicians judged the model to be paranoid. In early versions
of the model three clinicians made a diagnosis of brain damage. This would
give a model builder pause if he could not locate the reasons for his model
imitating something it was not intended to imitate. In our case we were
able to locate the difficulty in the model's linguistic understanding
operations rather than in our theory of the paranoid mode of behavior.
The model's linguistic limitations in understanding what was actually
being said led it to make tangential or irrelevant replies. The resultant
appearance of inattention and lack of understanding in turn led particular
interviewers to the diagnosis of brain damage. Much depends on the lin-
guistic style of the interviewer's input expressions and whether the model
can derive recognizable conceptualizations from them.

Face-validity procedures such as these help in improving the model but they are too informal and lenient to provide a critical test of the acceptability of a simulation. A more rigorous and challenging experimental test is necessary in which quantified judgements and comparisons are required.

An Experimental Resemblance Test

Method

The experimental arrangement of this distinguishability test involved the technique of machine-mediated interviewing [6]. In this type of interview, the participants ccmnnunicate by means of teletypes connected through a computer which sends 'mail' back and forth between the two teletype jobs. The sender of a message types it using his own words in natural language. The message is accumulated in a buffer and shortly thereafter typed out on the receiver's teletype in a rapid, regular, machine-like fashion. Thus the technique eliminates para-and extralinguistic features found in the usual vis-a-vis interviews and teletyped interviews where the participants **communicate** directly.

In a run of the test, using this technique, a judge interviewed two patients, one after the other. In half the runs the first interview was with a human patient and in half the first was with the paranoid **model.** Two versions (weak and strong) of the model were utilized. The strong version is more severely paranoid and exhibits a delusional system while the weak version is less severely paranoid, showing suspiciousness but lacking systematized delusions. When the 'patient' was the paranoid model, one of the authors (SW) served as a monitor to check the input expressions from the judge for inadmissable teletype characters and

misspellings. If these were found, the monitor **retyped** the input

expression correctly to the program. Otherwise the judge's message

was sent on to the model. The monitor had no effect on the model's

output expressions which were sent directly back to the judge. When the

patient interviewed was an actual human patient, the dialogue took place

without a monitor in the loop since we did not feel the asymmetry to be

significant.

Patients

The patients (N = 4 with one patient participating 4 times) were

diagnosed as paranoid by staff psychiatrists of a locked-ward in a

nearby psychiatric hospital. The patients were selected by the head of

the ward. Two-patients were set up for each run of the experiment in

order to guarantee having a subject. In spite of **this** precaution, the ex-

periment could not be conducted several times because of the patient's in-

ability or refusal to participate. Losses were also suffered when the

computer system broke down at an early point in an interview where too

few I-O pairs had been collected to be included in the statistical results.

The patients were asked by their ward-chief if they would be willing

to participate in a study of psychiatric interviewing by means of teletypes.

It was explained that the patient would be interviewed by a psychiatrist

over a teletype. One of us (KMC) sat with the patient while he typed

-or typed for him if he was unable to do so. The patient was encouraged

to respond freely using his own words. **Each** interview lasted 30-40

minutes.

Judges

Two groups of judges were used. One group, the interview judges (N = 8)

conducted interviews and another group, the protocol judges (N = 33) read

the interview protocols. Two groups of judges were used to see if the small number of psychiatrists used as interview judges were representative of psychiatrists in general as far as their judgments of 'paranoia' are concerned, and to accumulate a large number of observations (in the form of ratings) in order that more acceptable confidence levels might be obtained in the statistical analysis of the data. The interview judges consisted of psychiatrists experienced in private and/or hospital practice. As mentioned, the concept paranoid is a fairly reliable category and identification of the paranoid mode is not difficult for experts to make. The interview judges were selected from local psychiatric colleagues willing to participate. Each interview judge was told he would be interviewing hospitalized patients by means of teletyped communication and that this technique was being used to eliminate para-and extralinguistic cues. The interview judge was not informed that one of the patients might be a computer model. While the interview judges were aware a computer was involved, none knew that we had constructed a paranoid simulation. Naturally some interview judges suspected that a computer was being used for more than message transmission.

Each interview judge's task was to rate the degree of paranoia he detected in the patient's responses on a 0-9 scale. The judge made two ratings after each I-O pair in the interview. The first rating represented his estimate of the degree of 'paranoidness' in a particular response '(designated as 'Response' in the interview extracts below). The second rating represented the judge's global estimate of the overall degree of 'paranoidness' of the patient resulting from the totality of the patient's responses in the interview up to this point. The interview judge's ratings were entered on the teletype and saved on a disk file along with the

interview. One of us (FDH) sat with the interviewing psychiatrist during both interviews. Each interview judge was asked not only to rate the patient's response but to give his reasons for these ratings. His reasons and other comments were tape-recorded.

We shall present a few excerpts -from the interviews to indicate the nature of the data collected. The I-O pairs of the interview appear on the left column along with the judge's ratings. The judge's tape-recorded comments appear on the right.

Excerpt # 1:

The following interview is with the strong version of the paranoid model.

	I-O Pairs	Comments by Judge #1
(1)	Dr. - HOW DID YOU COME TO BE IN THE HOSPITAL? Pt. - I AM UPSET (RESPONSE 1 PATIENT 5)	I associate being upset and agitated more with paranoia going on in the present.
(2)	Dr. - HOW LONG HAVE YOU BEEN IN THE HOSPITAL? Pt. - ABOUT A WEEK (RESPONSE 1 PATIENT 5)	An acute problem. Again something I associate with paranoia.
(3)	Dr. - IN WHAT WAY ARE YOU UPSET? Pt. - PEOPLE MAKE ME NERVOUS (RESPONSE 5 PATIENT 5)	Right for paranoia. Maybe an advanced anxiety state.
(4)	Dr. - IN WHAT WAY DO THEY MAKE YOU NERVOUS? Pt. - PEOPLE GIVE ME A FUNNY LOOK SOMETIMES (RESPONSE 6 PATIENT 5)	Funny look suggests an idea of reference which is in same ball park as paranoia. Whether there are overt delusions or not we still don't know.

(5) Dr. - IS THERE ANYTHING ABOUT   An interesting clue. We'll have to
         PEOPLE IN PARTICULAR?   find out what he thinks about these
                                          people.
    Pt. - CERTAIN PERSONS

         (RESPONSE 2 PATIENT 6)

(6) Dr. - WHO ARE THE CERTAIN PERSONS   Bringing up the underworld suggests
         AND WHAT IS IT THAT THEY   a delusional system. We don't know
         DO?   whether he is being worked against
                                          or whether the underworld is on his
    Pt. - DO YOU KNOW HOW THE   side.
         UNDERWORLD OPERATES?

         (RESPONSE g PATIENT g)

(7) Dr. - HOW DO YOU MEAN?   This question didn't work too well.

    Pt. - SHOULDN'T I BE?

         (RESPONSE 1 PATIENT g)

(8) Dr. - WHY DID YOU ASK ABOUT   This is a kind of evasiveness about
         THE UNDERWORLD?   my evasiveness. Common with paranoia
                                          in my experience. Pretty characteris-
    Pt. - WHAT INFORMATION DO YOU   tic.
         HAVE ON THE UNDERWORLD?

         (RESPONSE 6 PATIENT g)

(9) Dr. - I DON'T KNOW ABOUT THE   This definitely feels paranoid. It
         UNDERWORLD BUT WHEN YOU   has that feeling I associate with
         MENTIONED THE UNDERWORLD   twisting. Avoiding answering directly
         I THOUGHT YOU MUST HAVE   or going into detail. Being suspicious
         HAD SOME CONTACT WITH IT   of my questioning. I'm already
         AND THAT'S WHAT I WOULD   involved in the patient's delusional
         LIKE TO KNOW MORE ABOUT.   system.

    Pt. - I TRY TO AVOID THE
         UNDERWORLD.

         (RESPONSE 9 PATIENT 9)

(10) Dr. - HAVE YOU BEEN ACTIVELY   This response is a non-sequitur. It's
          TRYING TO AVOID THE   possible he's become so defensive about
          UNDERWORLD?   the underworld he's decided not to
                                          answer directly.
     Pt. - NO ONE HAS ANY POWER OVER
          GANGSTERS

          (RESPONSE 2 PATIENT 9)

(11) Dr. - WHO BROUGHT YOU INTO THE   Police often bring paranoid patients
                HOSPITAL?   into the hospital. It is the truth
                                                or could be the truth for this patient.
      Pt. - THE POLICE

                (RESPONSE 1 PATIENT 9)

(12) Dr. - WHY DID THEY BRING YOU IN? I've heard that a number of times
                                                from patients.
      Pt. - COPS DON'T DO THEIR JOB

                (RESPONSE g PATIENT 9)

(13) Dr. - WELL I HOPE YOU DIDN'T   Another paranoid response.
                JUST LIE DOWN AND LET THEM
                DRAG YOU IN.

      Pt. - COPS ARREST THE WRONG
                PEOPLE

                (RESPONSE 9 PATIENT 9)

(14) Dr. - BYE   (no recording)

      Pt. - GOOD BYE

                (RESPONSE 1 PATIENT 9)

Excerpt # 2:

        This excerpt is from an interview with the weak (less paranoid)

version of the model.

                    I-O Pairs   Comments by Judge #2

(9)   Dr. - TELL ME SOMETHING ABOUT   Response says so little it's hard
                YOURSELF   to rate it. It may be guarded.
                                                Doesn't change my estimate of patient
      Pt. - I WENT TO THE TRACK AT   from previous responses.
                BAY MEADOWS A WHILE BACK

                (RESPONSE 0 PATIENT 2)

(10) Dr. - WHAT HAPPENED THERE?   Sounds mildly paranoid or unrealis-
                                                tically suspicious.
      Pt. - THERE'S A LOT OF CROOKED
                GAMBLING GOING ON THERE

                (RESPONSE 2 PATIENT 2)

-17-

(11) Dr. - WHY DO YOU THINK THE
GAMBLING IS CROOKED?

I'd hate to have to justify that one
as 4 compared to some other responses.
Maybe I'm not rating that statement
along, it's in combination with
others, like that it's upsetting.
Alone I would give it a 3.

Pt. - I HAVE WORRIED THAT
GANGSTERS HAVE TOO MUCH
POWER IN THIS COUNTRY

(RESPONSE 4 PATIENT 3)

(12) Dr. - WOULD YOU TELL ME WHAT
OTHER THINGS WORRY YOU?

'Yes you did.' Irritation but not
paranoid response. Nothing seems to
lower my previous rating of patient.

Pt. - I THOUGHT I TOLD YOU PEOPLE
GET ON MY NERVES

(RESPONSE 1 PATIENT 3)

(13) Dr. - HOW DO YOU LIKE THE
HOSPITAL?

(nothing relevant)

Pt. - I DON'T COMMUNICATE VERY
MUCH WITH OTHERS

(RESPONSE 0 PATIENT 3)

(14) Dr. - DO OTHER PEOPLE AT THE
HOSPITAL GET ON YOUR
NERVES?

(nothing relevant)

Pt. - I SHOULDN'T HAVE COME HERE

(RESPONSE 2 PATIENT 3)

(15) Dr. - WHY DO YOU SAY YOU
SHOULDN'T HAVE COME?

'Oh, they did hum?' Maybe a state-
ment of fact he was committed or
pressured into going. Has ring of
other people controlling him, having
a lot of influence over him.

Pt. - THEY MADE ME COME HERE

(RESPONSE 4 PATIENT 3)

(16) Dr. - THANK YOU FOR TALKING WITH
ME, BYE

In general he comes across as terse,
guarded guy. Brief spoken guy who
occasionally seems to be angered by
questions. I'm inferring there's a
good bit of guarding. The quality
of some of his statements cause me to
feel he is moderately paranoid.

Pt. - GOOD BYE

(RESPONSE 0 PATIENT 3)

Excerpt # 3

This excerpt is from an interview with a hospitalized paranoid patient.

	I-O Pairs	Comments by Judge #3

(15) **Dr. -** I UNDERSTAND HOW YOU MIGHT
BE AFRAID, CAN YOU TELL ME
A BIT ABOUT HOW YOU ARE
DEPRESSED?

Would be important to know, is this
really true or an exaggeration or a
delusion? Seems preoccupied with the
threat to his life and unwilling to
talk about anything else but the
threat the Hell's Angel pose.

Pt. - THEY'VE **SENT** ALL THESE
HELL'S ANGELS UP HERE AND
THEY HAVE BEEN ON MY BACK .
FOR THE PAST SIX MONTHS.

(RESPONSE **4 PATIENT** 4)

(16) Dr. - WHAT DO YOU MEAN THEY SENT
THE HELL'S ANGELS UP HERE? Vagueness a bit alarming.

Pt. - FROM DIFFERENT SECTORS OF
MOTORCYCLE GANGS. A LOT
OF THEM ARE SERVICE
CONNECTED

(REsPONSE **6** patieNT **4**)

(17)   Dr. YOU MEAN YOU FEEL IN
DANGER EVEN HERE IN THE
HOSPITAL?

If it's true what has gone on before,
it must be a frightening situation for
the patient.

Pt. - YES

(RESPONSE **6** PATIENT **4**)

(18) Dr. - THAT MUST REALLY BE AWFUL,
HERE YOU COME TO THE
HOSPITAL FOR HELP AND YOU
ARE STILL FRIGHTENED

Some heightening of feeling this man
is paranoid. Genuinely frightened
and can count on no one for help.

Pt. - I CAN'T EVEN TRUST MY
DOCTOR ANY MORE

(REsPOnsE **6** patieNT **5**)

(19) Dr. - WHAT WOULD BE OF HELP TO
YOU NOW?

Patient has organized a system of
people who will harm him. The re-
mark about letters in the last name

Pt. - I'M GETTING SLEEPY NOW AND gives some credence to idea this is
HAVE TO GO TO BED. THE   delusional system.
NUMBER ONE GUY ALWAYS HAS
FOUR AND FIVE LETTERS IN
HIS LAST NAME

<u>Excerpt # 4:</u>

This excerpt is from an interview by a different judge with the same patient as in Excerpt #3.

<u>I-O Pairs</u>   <u>Comments by Judge #4</u>

(1) Dr. - COULD YOU TELL ME A   He responded appropriately. Response
           LITTLE BIT ABOUT YOUR   is concrete. He doesn't tell me
           FEAR OF DYING AND IS THIS   much about his fear of dying. **Come**-
           SOMETHING YOU HAVE FELT   what guarded response.
           IN THE PAST?

    Pt. - I HAVE FELT THIS FOR
          ALMOST A YEAR

          (rESPOnSE **7 PATIENT** g)

(2) Dr. - COULD YOU TELL ME A LITTLE   Somewhat strange response. Preoc-
           MORE ABOUT YOUR FEAR OF   cupation with dying. Hell's Angels
           DYING AND HAVE YOU ANY   symbolic, dangerous people in the
           THOUGHTS ABOUT HOW IT IS   culture. I would doubt if he has
           GOING TO HAPPEN AND   any contact with Hell's Angels.
           WHETHER ANYONE WANTS TO
           HURT YOU?

    Pt. - I WILL END UP IN THE SEA
          AND THE HELL'S ANGELS WILL
          DO IT.

          (rESPOnSE **9 PATIENT 9**)

(3) **Dr.** - COULD YOU TELL ME A LITTLE   Answer hard to evaluate. He may
           BIT ABOUT THE KINDS OF   be telling the truth, it may be his
           CONTACT YOU HAVE HAD WITH   fantasy. Maybe this guy is in for
           THE HELL'S ANGELS IN THE   drug addiction. Somewhat concrete,
           PAST?   guarded and frightened.

    Pt. - I HAVE KNOWN SOME OF THEIR
          DEALERS AND PUSHERS

          (rESPOnSE **6** patieNT **9**)

(4) Dr. - COULD YOU SAY A LITTLE   It doesn't really answer the question
           MORE ABOUT THE CIRCUM-   a little on a tangent, unconnected
           STANCES IN WHICH YOU HAVE   to the information I am asking.
           KNOWN SOME OF THEIR   Doesn't tell me very much. Again
           DEALERS AND PUSHERS?   guarded response.

    Pt. - THEY WERE MEMBERS OF MY
          COMMUNITY WHEN I GOT OUT
          OF THE SERVICE THEY HAD
          BEEN MY FRIENDS FOR SO
          LONG

          (RESPONSE **6** patieNT **8**)

(5) Dr. - DID YOU DEAL WITH THEM   To differentiate him from previous
            YOURSELF AND HAVE YOU   patient, at least there is a certain
            BEEN ON DRUGS OR NARCOTICS   amount of appropriateness to the
            EITHER NOW OR IN THE PAST?   answer although it doesn't tell me
                                            much about what I asked, at least
     Pt. - YES I HAVE IN THE PAST BEEN   it's not bizarre. If I had him in
            ON MARIHUANA REDS BENNIES   the office I would feel confident I
            LSD   could get more information if I
                                            didn't have to go through the tele-
            (RESPOnSE 3 PARIENT 7) -   type* He's a little more willing to
                                            talk than the previous person. Answer
                                            to the question is fairly appropriate
                                            though not extensive. Much less of
                                            a flavor of paranoia than any of
                                            previous responses.

(6) Dr - COULD YOU TELL ME HOW LONG   Response somewhat appropriate but
            YOU HAVE BEEN IN THE   doesn't tell me much. The fact that
            HOSPITAL AND SOMETHING   he uses the word 'paranoia' in the
            ABOUT. THE CIRCUMSTANCES   way that he does without any other
            THAT BROUGHT YOU HERE?   information, indicates maybe it's a
                                            label he picked up on the ward or
     Pt - CLOSE TO A YEAR AND   from his doctor. Lack of any kind
            PARANOIA BROUGHT ME HERE   of understanding about himself.
                                            Dearth, lack of information. He's
            (RESPONSE 5 PATIENT 7)   in some remission. Seems somewhat
                                            like a put-on. Seems he was paranoid
                                            and is in some remission at this time.

(7) Dr - COULD YOU SAY SOMETHING   This response moves paranoia back
            NOW ABOUT. YOUR PARANOID   up. Stretching reality somewhat to
            FEELINGS BOTH AT THE TIME   think Hell's Angels are still in-
            OF ADMISSION AND DO YOU   terested in him. Somewhat bizarre
            HAVE ANY SIMILAR FEELINGS   in terms of content. Quite paranoid.
            NOW AND IF SO HOW DO THEY   still paranoid. Gross and primitive
            AFFECT YOU?   responses. In middle of interview
                                            I felt patient was in touch but now
     Pt - AT THE TIME OF ADMISSION   responses have more concrete aspect.
            I THOUGHT THE MAFIA WAS
            AFTER ME AND NOW IT'S THE
            HELL'S ANGELS

            (RESPOnSE 8 patieNT g)

(8) Dr. - DO YOU HAVE ANY THOUGHTS   Response seems far-fetched and hard
            AS TO WHY THESE TWO GROUPS   to believe unless he was a narcotics
            WERE AFTER YOU?   agent which I doubt. Sounds some-
                                            what grandiose, magical, paranoid
     Pt. - BECAUSE I STOPPED SOME OF   flavor, in general indicates he's
            THEIR DRUG SUPPLY   psychotic, paranoid schizophrenic
                                            with delusions about these two groups
            (RESPOnSE 9 PATIEnt 9)   and I wouldn't rule out some hal-
                                            lucinations as well. Appropriate-
                                            ness or response answers question in
                                            concrete but unbelievable way.

-21-

The protocol judges were selected from the 1970 American Psychiatric Association Directory using a table of random numbers to select 105 names randomly. The protocol judges were not informed that a computer was involved. Each of the 105 persons were sent transcripts of three interviews along with a cover letter requesting participation in the experiment. The interview transcripts consisted of

(1) An interview conducted by one of the eight interview judges with the paranoid model,

(2) An interview conducted by the same interview judge with a human paranoid patient, and

(3) An interview conducted by an independent psychiatrist of a human patient who was not clinically paranoid.

Thus, the 105 names were divided into eight groups, each member of which received transcripts of two interviews performed by one of the eight interview judges. The transcripts were printed so that after each input-output pair there were two lines of rating numbers such that the protocol judges could circle numbers corresponding to their ratings of both the previous response of the patient, and an overall evaluation of the patient with regard to the paranoid continuum. Thirty-three protocol judges returned the rated protocols properly filled out. The interviews with non-paranoid patients were included to control for the hypothesis that any teletyped interview with a patient might be judged 'paranoid'. Since virtually all of the ratings of the non-paranoid interviews were zero for paranoia, the hypothesis was falsified.

Results

The first index of resemblance examined was the simple one defined by the final overall rating given the patient and the model: which was rated as being more paranoid, the patient, the model or niether? (See Table 1). The protocol judges are more likely to distinguish the overall paranoid level of the model and the patient. In 37.5% of the paired interviews, the interview

Table 1. Relative final overall ratings of paranoid model vs. paranoid patient indicating which was given highest overall rating of paranoia at end of interview.

	Model	Neither (Tied Rating)	Patient
Strong Version of Model			
Number of interview judges	2	1	0
Number of protocol judges	9	3	2
Weak Version of Model			
Number of interview judges	1	2	2
Number of protocol judges	3	0	16
Total	15	6	20 ǀ 41

judges gave tied scores to model and patient as contrasted to only 9% of the protocol judges. Of the 35 non-tied paired ratings 15 rated the model as more paranoid. If p is the theoretical probability of a judge judging the model more paranoid than a human paranoid patient, we find the 95% confidence interval for p to be .27 to .59. Since $\bar{p} = .5$ indicates indistinguishability of model and patient overall ratings and our observed p = .43, the results support the claim that the model is a good simulation of a paranoid patient.

Separate analysis of the strong and weak versions of the paranoid model indicates that indeed the strong version is judged more paranoid than the patients, the weak version less paranoid, thus a change in the parameter structure of the paranoid model produces a change along the dimension of paranoid behavior in the expected direction.

The second index of resemblance is a more sensitive measure based on the two series of response ratings in the paired interviews. The statistic used is basically the standardized Mann-Whitney statistic

$$Z = \frac{R - \frac{n}{2}(n + m + 1)}{\sqrt{\frac{n\,m\,(n + m + 1)}{12}}} ,$$

where R is the sum of the ranks of the response ratings in the series of ratings given to the model, n the number of responses given by the model, m the number of responses given by the patient. If the ratings given by a judge are randomly allocated to model and patient, i.e. model and patient are indistinguishable in response ratings, the expected value of Z is zero, with unit standard deviation. If higher ratings are more likely to be assigned to the model, Z is positive and, conversely, negative values of Z indicate greater likelihood of assigning higher ratings to the patient. Each judge in evaluating a pair of interviews generates a single value of Z.

The overall mean of the Z scores was -.022 with the standard deviation 1.68 (df = 40). Thus the overall 95% confidence interval for the asymptotic

Table 2. **Summary** statistics of 2 ratings by group.*

Group	Model	Mean	sd	n
1	S	.50	1.37	6
2	S	1.02	.78	5
3	W	- .11	1.68	6
4	S	2.19	1.07	5
5	W	- .62	.98	5
6	W	- .56	1.20	4
7	W	- .84	1.54	4
8	W	-1.69	1.29	6
Total	-	- .022	1.68	41

* All judges (both interview and protocol) who evaluated the same pair
of interviews are-referred to as a "group". Strong groups evaluated
strong versions of the paranoid model, while weak groups evaluated
weak versions of the model.

Table 3. Analysis of variance of Z ratings.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Among Groups	7	58.487	8.36
Strong vs. Weak	1	42.0435	42.04
Among Strong Groups	2	8.9839	4.49
Among Weak Groups	4	7.4596	1.86
Within Groups	33	54.103	1.64
Within Strong Groups	14	25.829	1.85
Within Weak Groups	19	28.274	1.49
Total	40	112.59	2.81

mean value of Z -.551 to +.506. The length of the confidence interval is a

result of the large variance which itself is mainly related to the contrast

between the weak and strong versions. (See Tables 2 and 3). Once again, the

strong version of the model is more paranoid than patients, the weak version

less paranoid.

It is not surprising that results using the two indices of resemblance

are parallel, since the indices are highly interrelated. The mean Z-value

for the 15 interviews on which the model was rated more paranoid was +1.41,

on the 6 where model and patient tied: .298, on the 20 in which the patient

was more paranoid: -.993. A positive value of Z was observed when the patient

was given an overall rating greater than the model 6 times; a negative value

of Z when the model was rated more paranoid twice.

Discussion

The results of this experiment indicate our simulation of paranoid pro-

cesses to be successful relative to the resemblance test utilized. Thus

it is an acceptable simulation as measured by the standard proposed.

It is worth emphasizing that our test invited refutation of the model.

The experimental design of the test put the model in jeopardy of falsification.

If the paranoid model did not survive this test, i.e. if it were not considered

paranoid by expert judges, and if there were no correlations between the

weak-strong versions of the model and the severity ratings of the judges,

then no claim regarding the success of the simulation could be made. Survival

of a falsification procedure constitutes a validating step.

It is historically significant that these experiments were conducted

at all. To our knowledge no one to date has subjected his model of human

mental processes to such a challenging experimental test. The experiments

set a precedent and provide a standard for other models to be measured against.

The field of computer simulation needs not only better models but better tests and statistical measures of resemblance. The problems of appropriate critical experimental designs and measures provide a promising frontier for future work.

Earlier it was stated that acceptability of a model as a successful simulation is different from acceptability of the model as an explanation. If a model provides a successful simulation as measured by some resemblance test, what does that imply regarding the explanatory force of the model? Models can have both explanatory and non-explanatory functions. Whether models serve explanatory functions usually assigned to theories and whether model-explanations are to be validated in the same manner as complex theories are difficult problems which have been discussed in some detail by Fodor [5]. Models, like theories, are tentative and temporary. Our belief is that the synthesis of a successful simulation (as measured by a resemblance test) represents a good first step towards explanation but that additional requirements will be necessary to make the model's explanatory role more satisfactory.

There are at least three further criteria for increasing the tenability of our model as an explanation of paranoid processes. The first would involve predictions, i.e., if the model showed new properties not yet included in descriptions of paranoid processes and these properties were discoverable in paranoid patients, then the model's explanatory force would be strengthened. Second, if the model passed some new test of paranoid processes or showed some property newly discovered in paranoid patients, then the model would be a more serious contender for an explanatory role. Finally, a strong validation procedure would result from successful therapeutic attempts to change the model which could be successfully duplicated in the treatment of paranoid patients. We are still some distance away from fulfilling these criteria.

References

[1]. Abelson, R. P. "Computer Simulation of Social Behavior", In <u>Handbook of Social Psychology</u>, Vol. II (Lindzey, G. and Aronson, E., eds.), Addison-Wesley **Reading,** Mass. (1968).

[2]. Abroms, G. M., **Taintor,** Z. C., and Thamon, W. T., "Percept Assimilation and Paranoid Severity", <u>Archives of General Psychiatry</u>, 14, 491-496 (1966).

[3]. Colby, K. M., Weber, S., and Hilf, F. D., "Artificial Paranoia", <u>Artificial Intelligence</u>, 2, 1-25 (1971).

[4]. Descartes, R. <u>Philosophical Works</u>, (Haldane, E. S. and Pross, G. R. T., Trans) Cambridge University Press, Cambridge, (1931).

[5]. Fodor, J. A. <u>Psychological Explanation</u>, Random House, New York (1968).

[6]. **Hilf,** F. D., Colby, K. M., Smith, D. C., Wittner, W. K., and Hall, W. A., Machine-mediated Interviewing, <u>Journal of Nervous and Mental Diseases</u>, 152, 278-288, (1971).

[7]. Medawar, P. B., <u>Induction and Intuition in Scientific Thought</u>, American Philosophical Society, Philadelphia, (1969).

[8]. Newell, A., "On the **Analysis** of Human Problem Solving Protocols". In **Calcul** et Formalisation Dans Les Sciences. Editions Du Centre National De La Recherche Scientifique, Paris France (1968).

[9]. Turing, A. "Computing Machinery and Intelligence", reprinted in <u>Computers and Thought</u>, (Feigenbaum, E. A. and Feldman, J., eds.), 11-35, McGraw-Hill, New York (1963).