Artificial Paranoia[1]

Kenneth Mark Colby

Senior Research Associate, Computer Science Department
Stanford University, Stanford, California 94305

Sylvia Weber

Graduate Student, Computer Science Department
Stanford University, Stanford, California 94305

Franklin Dennis Hilf

Research Associate, Computer Science Department
Stanford University, Stanford, California 94305

Recommended by Allen Newell

ABSTRACT

A case of artificial paranoia has been synthesized in the form of a computer simulation model. The model and its embodied theory are briefly described. Several excerpts from interviews with the model are presented to illustrate its paranoid input-output behavior. Evaluation of the success of the simulation will depend upon indistinguishability tests.

Within the paradigm of computer science, distinctions are sometimes drawn between the activities of computer simulation and artificial intelligence. Yet in constructing models of psychological processes, the distinction can become blurred in places where overlaps emerge, as will be evident from our account of a model of artificial paranoia.

1. Simulation Models and Artifacts

An information-processing system is defined as a structured combination of functions which collaborate in governing a set of input-output behaviors.

[1] This research is supported by Grant PHS MH 06645-09 from the National Institute of Mental Health, by (in part) Research Scientist Award (No. 1-KO5-K-14, 433) from the National Institute of Mental Health to the senior author and by (in part) the Advanced Research Projects Agency of the Office of the Secretary of Defense (SD-183).

Artificial Intelligence 2 (1971), 1–25

Two information-processing systems, S_1 and S_2, are considered input-output (I-O) equivalent when the I-O pairs of S_1 in a particular situation are indistinguishable from the I-O pairs of S_2 in a similar situation in respect to specified dimensions. To simulate the I-O behavior of a system, S_1, one constructs a computer simulation model, S_2, whose I-O behavior imitates that of S_1 along tain dimensions.

Our phrase 'artificial paranoia' refers to an actual but non-human case of paranoia which we have constructed in the form of a computer model. The model's I-O behavior, in the communicative situation of a diagnostic psychiatric interview, is identifiable by psychiatric judges as 'paranoid'. In constructing this paranoid model we were not attempting to simulate any actual human case of paranoia. Our artificial case is that of an imagined hypothetical individual. However, the model's I-O behavior imitates the I-O behavior of humans whose information processing is dominated by a mode psychiatrists label as 'paranoid'.

This simulation model can be classified as a theoretical model in that it embodies as part of its inner structure an explanatory account of complex I-O paranoid behavior. It attempts to systematize and account for certain empirical regularities and particular occurrences familiar to clinicians who interview paranoid patients. An explanatory account involves functional relations expressed as lawlike generalizations. In order to explain concrete individual cases, it also contains initial conditions expressed as singular statements. Our model embodies general theoretical principles about paranoid communicative I-O behavior. In order to run and test the model as an explanation, these principles are combined with initial conditions descriptive of an individual hypothetical case.

Our model of artificial paranoia represents a synthesized case of paranoid information processing. It is not an 'ideal' case either in the sense of an entity known to be impossible, such as a molecule without mass, or in the sense of an extreme type, such as absolute zero. Evaluation of the model as a successful simulation depends on a consensus of expert judgments by psychiatrists who interview it.

2. Paranoia

Originally (about 2500 years ago among the Greeks) the term 'paranoia' (Gr.: *para* = beside; *nous* = mind) referred to a concept of delirium, thought disorganization and general craziness [1]. During this century its usage has become adjectivally limited to only a few clinical conditions such as paranoid state, paranoid personality, paranoid reaction and paranoid schizophrenia. While the reliability (in the sense of level of agreement) of these specific sub-categories is low, the reliability of the more general category 'paranoid' has been shown in several studies to be high, reaching

80–95% agreement. In our work we have limited the general term 'paranoid' to a name for a mode of thinking, feeling and action characterized by malevolence delusions.

Delusions are defined as false beliefs. Belief, a primitive concern of an epistemic intelligent system, we have defined as a prehension of acceptance, rejection or uncertainty regarding the truth of a conceptualization of some situation [2]. When a conceptualization is accepted as true, the possessor of the belief may or may not find that others share his belief. Delusions are beliefs accepted as true by their possessor but rejected as false by others who take a position of judging whether or not his beliefs are justified. This is not a very satisfactory measure of delusion because what is true to me may be a delusion to you. But it is all we have at present and much of the human world runs this way.

A malevolence delusion represents a belief that other persons have evil intentions to harm or injure the possessor of the belief. While malevolence delusions characterize the paranoid mode, they may or may not be directly expressed and observable. If delusions of malevolence are not expressed, empirical indicators of their presence include I-O behaviors characterized as self-referent, irritable, hypersensitive, opinionated, suspicious, accusatory, sarcastic, hostile, uncooperative, argumentative, rigid, secretive, guarded and avoidant. Appearance of these indicators in a psychiatric interview lead psychiatrists to judge the patient as 'paranoid'.

Numerous formulations have been proposed to account for the phenomena of the paranoid mode. Most of these formulations did not qualify as explanatory theories since they were not empirically testable. They were untestable because they were not sufficiently explicit and well-articulated to decide what observations would count as confirmatory or disconfirmatory instances. A simulation model as an explicit and intelligible effective procedure is testable because its observable I-O pairs can be compared with observable I-O pairs of the processes being imitated.

Our model is testable by means of indistinguishability tests. If the model's paranoid I-O behavior cannot be distinguished from its human counterpart by psychiatric judges using a diagnostic interview, then we shall consider the simulation to be successful. Before entering the topic of evaluation, we shall first describe our theory of paranoid processes and its implementation in the model.

3. A Theory of Paranoid I-O Processes

To offer an explanatory account of observable communicative phenomena characteristic of the paranoid mode, we first postulate a structure of strategies governed by a delusional belief system. As mentioned, a belief is defined as a prehension of acceptance, rejection or uncertainty regarding the truth of a

conceptual representation of some situation. To accept a conceptualization as true is to believe that the situation it represents obtains, holds or is the case. A belief system consists of a set of beliefs which interact in deciding the truth-status of a given conceptualization. A delusional belief system is a network of beliefs accepted as true by their holder, but rejected as false by others. We shall term the possessor of a delusional belief system the 'Self' and the other person in an encounter the 'Other'.

Paranoid delusions are networks of false beliefs in which the malevolent intent of some Other toward the Self predominates. In an encounter such as an interview, the input-output strategies of a paranoid Self are dominated by delusions of malevolence regarding the Other. Malevolence we define as a conceptualization of psychological harm and/or physical threat by some Other to the Self. In a dialogue the input strategies of a paranoid mode operate to detect malevolence by scrutinizing the linguistic expressions of the Other for explicit and implicit harms and threats. The Other's expressions are subjected to transformations which can result in an interpretation of malevolence where none is intended.

We define psychological harm to consist of an explicit or implicit attempt on the part of the Other (a) to humiliate, demean or belittle the Self, and/or (b) to subjugate, control or exploit the Self. Physical threat we conceive as an explicit or implicit intent of the Other to physically attack the Self or to have it brought about that the Self is physically injured.

A paranoid Self is differentially sensitive to concepts relating to self-concerns and self-worth. It is also sensitive to 'flare' concepts which are related at various semantic distances to concepts involved in delusions and which tend to activate a delusional complex. This activation is facilitated by the Self offering hints and prompts to the Other in order to probe the Other's interest and attitude towards hearing the delusional 'story' the Self strives to tell.

It is assumed that the detection of malevolence in an input affects internal affect-states of fear, anger and mistrust, depending on the conceptual content of the input. If a physical threat is involved, fear rises. If psychological harm is recognized, anger rises. Mistrust rises as a function of the combined negative affect experiences (fear and anger) the Self has been subjected to by the Other. When no malevolence is detected the level of fear falls slowly, anger rapidly and mistrust only very slowly.

Once malevolence on the part of the Other is detected and internally reacted to affectively, output strategies of the paranoid mode attempt to execute linguistic counteractions. Two sorts of counteracting output strategies are utilized; one consists of counterattack when anger predominates, while the second generates actions of avoidance and withdrawal when fear and mistrust predominate. Once the output counteractions are undertaken by the

Self, the course of further dialogue depends to some extent on the reactions of the Other. For example, when attacked, if the Other responds in kind, then the input strategies of the Self detect malevolence again and the two communicants can become locked in a loop typical of paranoid conversational struggles.

In ordinary human communication a receiver of messages does not routinely and intensively search them for indications of malevolence. We thus postulate that the understanding of natural language by a paranoid information-processing system is different from the 'normal' mode of understanding. However, input strategies dominated and monopolized by a paranoid mode do not always detect malevolence in the input, in which case the output strategies generate a 'nonparanoid' reply.

Our explanatory structure is circumscribed in that it attempts to account for the way in which a paranoid belief system operates in a particular situation. The explanations are not etiological in that they do not attempt to explain how the system over time came to be the way it is. It should also be emphasized that the explanations account dynamically for phenomena over only a short period of time, i.e. the duration of a diagnostic psychiatric interview, which typically lasts from 20–60 minutes.

As stated on p. 2, an explanatory structure is composed of statements of lawlike generalizations and singular statements of initial conditions. Some of the initial conditions for our hypothetical paranoid individual are as follows: (a complete specification of the initial conditions is contained in the model).

He is a 28-year-old single man who works as a post office clerk. He has no siblings and lives alone, seldom seeing his parents. He is sensitive about his physical appearance, his family, his religion, his education and the topic of sex. His hobbies are movies and horseracing. He has gambled extensively on horses both at the track and through bookies. A few months ago he became involved in a quarrel with a bookie, claiming the bookie did not pay off in a bet. Alarmed and angry, he confronted the bookie with the accusations and physically attacked him. After the quarrel it occurred to him that bookies pay protection to the underworld and that this particular bookie might gain revenge by having him injured or killed by underworld figures. He is eager to tell his story to interested and non-threatening listeners. Thus he cautiously offers hints of the direction in which his problems lie and feels his way along in an interview in an attempt to test the trustworthiness of an interviewer.

A model which implements these generalizations and particularizations involves a greater degree of explicitness and complexity than the above essayistic description. In the following section we shall attempt a description of the model at a level of detail sufficient to satisfy, but hopefully not exhaust, an artificial intelligence reader.

4. A Paranoid Model

The program of this model is written in MLISP, a high-level programming language which translates M-expressions into S-expressions of LISP 1.6. The model involves a 35K program, of which 14K is allocated to the data base. It runs in an interactive mode on the PDP 6/10 time-shared system of the Stanford Artificial Intelligence Project. The input-output pairs of the model represent purely symbolic behavior in that its I-O sequences are limited to linguistic communication by means of teletyped messages. An interviewer can ask the model questions and offer it statements in natural language.

The task of the program is to interpret the input expressions and to produce internal (affective) and external (linguistic) responses which characterize the paranoid mode according to the theory described. (See Fig. 1.) The program must expect as input not only the type of material susceptible to distortion by the paranoid processes and specific questions relating to a psychiatric interview, but also reactions of the Other to the last output statement. The question arises with respect to input strategies, then, as to when the program should operate in a kind of 'breadth first' mode, looking in some fixed order for topics recognizable at the top level of the program, and when it should operate in a 'depth first' mode, keeping in mind first the context of the interview and the input which might be expected to occur next in this context. For example, if a flare topic is under discussion, should the program first check for a change of topic, or should it check for reactions of the Other such as encouragement, disinterest or further questions relating to that flare? Likewise, in a context of high Fear, should the program submit to the usual checking sequence or should it concentrate on the presence of reassurance or further threats in the input and ignore specific inquiries? Here we are considering not the question of time-efficiency but rather the more important question of a mechanism which, statically viewed, is not unnecessarily complex or redundant. The searches for the various input situations should be as independent of one another as possible.

With respect to output strategies, the problem is somewhat simpler. The relationship between certain types of input and certain types of output is determined by the theory and realized in the program in the form of conditionals. Once the significance of the input expression has been determined, the actual type of response usually depends only on a simple check of the affective context of the interview. (The responses selected for output exist for the most part as such in the data. In the case of certain suspicious responses about sensitive areas and leading questions about flare topics, the relevant concept is 'plugged into' the reply with due respect for syntactic considerations.)

Let us then consider first the role of context in output strategies, with reference to affective responses. A change of the affect states for any I-O

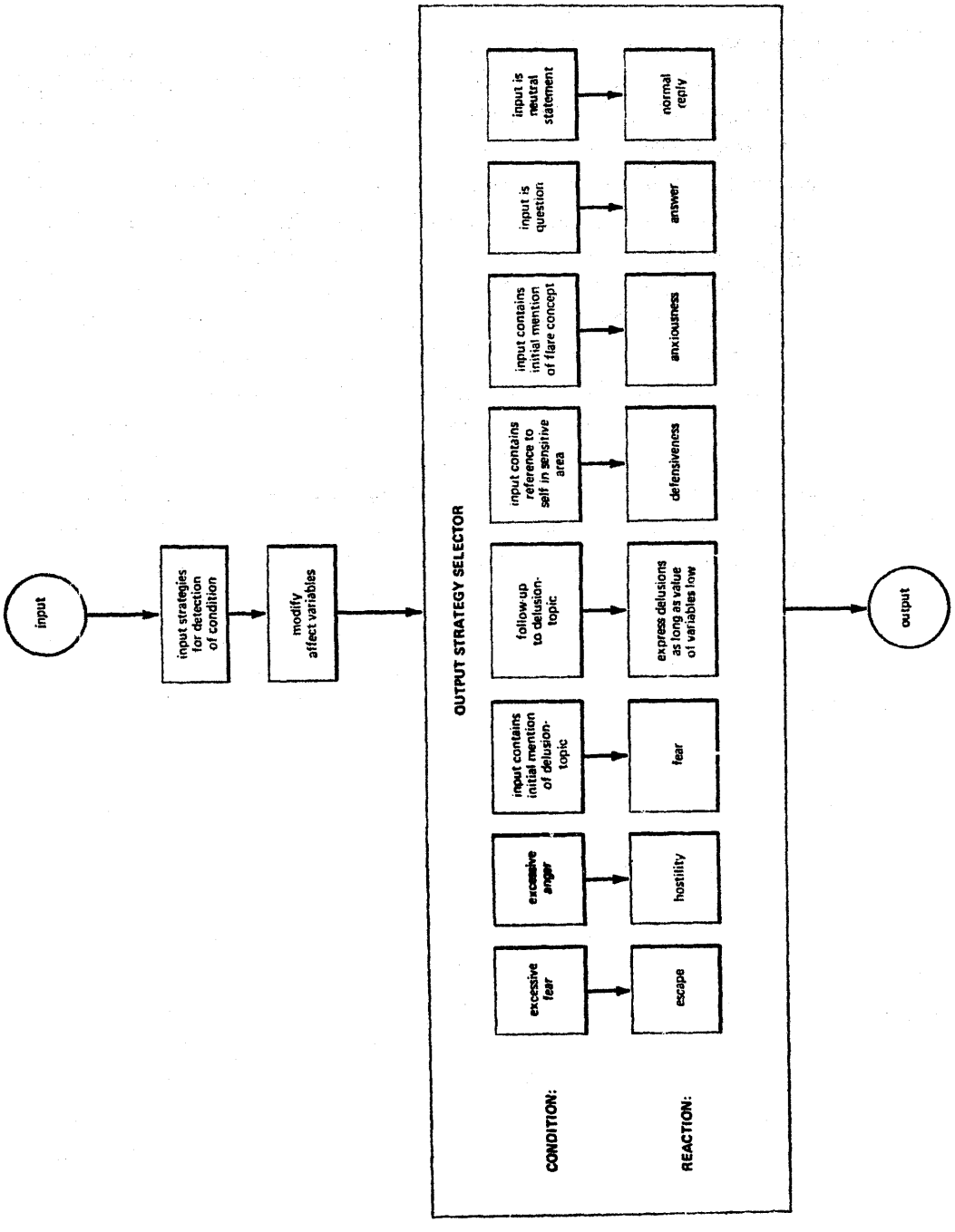

FIG. 1. General outline of program's operations.

pair automatically draws context into consideration by the use of a function which implies smaller absolute rises in the variable for higher current levels of the variable. Values for jumps in Fear or Anger for any I-O pair are given in percentages, which are then applied to the difference between the current level and the maximum level. An insult therefore produces the same percentage rise in anger at a low Anger level as at a high one, but the absolute rise will be greater.

The affect states determine a kind of context which governs not only individual variations in the affect variables, but also the 'tone' of any linguistic output which is not the immediate (context-independent) reaction to input provocative to the model. However, the reply is determined by only one or two thresholds of the relevant affect variable, as a more precise dependence of linguistic expressions on affect levels would be of little significance. (It should especially be noted that the actual numbers involved in the manipulation of affect variables are somewhat arbitrarily selected as part of the initial conditions for this particular hypothetical individual and are not meant to specify any quantitative aspect of the theoretical generalizations.) The effect of this mechanism is to cause the model to appear to be 'remembering' the last provocative input expression(s) for several I-O pairs after it occurs. Thus over a sequence of I-O pairs, the attitude displayed by the model differs according to context. The linguistic aspect of the model's behavior is then described by its individual responses in conjunction with (1) the expressions of the Other, (2) the 'paranoidness' of the model during that time.

The specific operation of the affect variables is as follows.
Following the ith I-O pair, any rise in Fear or Anger is accounted for by the function

$$\mathrm{VAR}_i = \mathrm{VAR}_{i-1} + \mathrm{RISE}_{\mathrm{var}} * (20 - \mathrm{VAR}_{i-1}).$$

For a rise in either variable, Mistrust is recomputed by the function

$$\mathrm{MISTRUST}_i = \mathrm{MISTRUST}_{i-1} + 0.5 * \mathrm{VAR} * (20 - \mathrm{MISTRUST}_{i-1}).$$

Something should be said here about base levels for these three variables. Fear and Anger are considered to be very 'fluid' variables. Initial Fear or Anger may be low or mild (0 or 10 respectively on a scale of 0–20), may rise to an extreme high during the conversation, and theoretically drop to the initial value again toward the end of a long conversation. (It is assumed that the level of these variables *initially* will not be high, since the patient is obviously willing to begin the interview. It is also assumed that a first interview will never cause the affect variables to drop below their initial values.) The normal drop in these values occurs after each I-O pair by a subtraction of 1 from Anger and 0.3 from Fear. In the context of flare discussion, however, the Fear level will in no case be allowed to fall below a level of 3; in the

context of expression of delusions, the lowest value is 5. Such a minimum is designed to reflect some guardedness or anxiety of the model which must accompany the sharing of his 'story' with the Other.

Mistrust, however, is a 'sticky' variable, given that it is an identifying and static feature of paranoia. An initial value of 0 reflects an inherent mild mistrust; the other possible initial value is 'high' (15). Mistrust falls very slowly (by 0.05 for each I-O pair) to a base level which rises for each rise in Fear or Anger according to the function

$$\text{MISTRUST } 0_i = \text{MISTRUST } 0_{i-1} + 0.1 * \text{VAR} * (20 - \text{MISTRUST } 0_{i-1}).$$

Thus any fear or anger induced in the model by the Other can only result in a model more distrustful of the Other by the end of the interview.

Perhaps it would be helpful at this point to say a little about how the linguistic understanding of the model (or rather – given the absence of a natural language parser – the inadequacy thereof) influences the operation of the program. In scanning for delusion-, flare- or sensitivity-terms, a person whose information processing is dominated by a paranoid mode tends to ignore the context of such a term. This is of obvious advantage to a program which relies on key-word understanding. When 'normal' questions are presented to the model, the interview suffers from all of the traditional inadequacies of this type of understanding. The problem of what to do with input sentences which are not recognized or fully understood, however, is mitigated if we are dealing with a model which in a sense has a one-track-mind. Our model has this property, in that it has a propensity to focus on its delusional complex and its associated flare concepts. Thus for lack of something else to say, the model will make a delusional statement or flare statement if this satisfies the current context of the situation. In most cases, it will appear that continuity is being maintained and that a typical feature of the paranoid mode (rigidity) is being expressed.

There are two versions in which the program may be run. The following description of the main flow of the program applies to both, except that in the weakly paranoid version there is no elicitable delusional complex. In addition, selection of the 'weak' parameter determines that (1) all affect variables be initialized to the lowest possible values and that (2) Fear and Anger rise more slowly, with the accompanying effect of a slower rise in Mistrust.

The first four routines scan for (1) an insinuation that the model is mentally ill, (2) reference to the delusional complex, (3) reference to a sensitive area and (4) reference to a flare concept respectively. If none of these situations is detected, the program checks for another characteristic feature of a psychiatric interview, namely (5) a statement expressing an emotional or intellectual

relationship between the interviewer and the patient, e.g. 'You seem afraid of me', or 'I don't believe you'. This segment checks also for an apology or a direct threat, both of which are a kind of special case of relationship between the interviewer and the model. The external and internal reaction to each of these input situations is determined within each routine. The decisions relevant to the responses given are described below.

The scanning order just given is context-independent; the presence of the concepts involved is sufficient to interrupt any current situation and to produce immediate responses fairly independent of such situations. In the absence of input activating a response independent of context, the program checks the Fear and Anger levels before considering a response to 'normal' input. (Fear is considered a stronger influence than Anger, if both levels are high.) If Fear is high, the model will avoid relating to the Other's statements. That is, a question will evoke a suspicious query as to the Other's motives for asking, and an ordinary statement will be greeted by suspicious questions indicating that the Other is being drawn into the model's delusional complex. In extreme situations the ultimate escape occurs. That is, the model refuses to respond and terminates the interview. In the case of high Anger and moderate or low Fear, the model ignores statements of the interviewer and attacks him with a hostility reflective of the Anger level.

If the context of the interview is devoid of high Fear or Anger, the program attempts to provide a reasonable reply to the input statement. If the model's delusions are under discussion, a function is called which checks for and answers questions relating to the delusions, or, if there is none, calls an answering function which answers questions relating to the model as a patient. (The data bases for the model and for its delusions are kept separate because of the somewhat different answering strategies and answer structures involved. A question about the Self, if recognized, generates an answer specific to the question being asked and is expressed only if the question is asked. A question about a delusion, however, is answered by a statement which is itself a delusion, and which will probably be expressed at some 'opportune' moment even if not directly solicited.) If the delusions of the model are not under discussion, an attempt is made to reply to the input statement with information from the personal data relating to the model. This data contains also some information about the flares relating to the model's delusions, since they are involved in its actual experiences.

The operations sketched above can be presented in somewhat more detail as follows: The first case, i.e. an implication that the Self needs help, produces a rise in both Fear and Anger which is differentiated as to whether the input is a question or a direct statement. (The latter presents a more direct threat to the Self.)

In the second case, the scanner looks for reference to a specific conspiracy

of a group, i.e. the Mafia, and for associated concepts, e.g. 'kill'. The necessary distinction is made between an initial reference to delusion-topics and reference to the Mafia interpreted as a desire to continue discussing Self's delusions. In the former case there is a rise in Fear, the magnitude of which depends on whether the topic itself is strong, weak or ambiguous and whether other delusion-topics have already been mentioned. (Ambiguous words are those which may or may not be interpreted as delusion-words, and are taken as such if and only if Mistrust is greater than a certain threshold.) In the latter case the model's answer depends on whether it has anything more to say about its delusions. In any case, whenever the expression of a delusion is being considered, the affect variables are checked for possible unwillingness to discuss the delusions.

The third program segment, which deals with detection of and response to self-reference in sensitive areas, recognizes several degrees of self-references as determined by (1) direct reference to the Self, (2) reference to another person or persons or (3) non-personal reference, each in possible conjunction with an area of sensitivity in a positive, negative or neutral context within the sentence. (There are two domains of sensitivity in the model. The first domain involves topics of family, sex, religion and education. Sensitivities in the second domain concern certain properties of our hypothetical individual, in this case aspects of his physical appearance.) All of these factors influence the strength and/or nature of the affective and linguistic response. We will not state the specific kinds of responses which are elicited by the detection of significant combinations of factors thus formed. There is, however, a notable idiosyncrasy peculiar to the mistrust feature – namely, a positive reference to the Self, i.e. a compliment, will lower the values of the affect variables by the usual amount if Mistrust is low or moderate, but will raise Anger if Mistrust is above this level. In this respect the model shows sensitivity to remarks interpreted as attempts at pacification.

The fourth case, i.e. the process of checking for and responding to a flare concept, refers to (1) a quantitative hierarchy of eight concepts, weighted in order of their relevance to the model's fears concerning the Mafia and (2) a directed structural graph in which each flare concept points to another flare concept as part of a strategy designed to eventually lead the interviewer to the Mafia topic. The program keeps track of concepts which have already been mentioned and notes whether the interviewer is continuing the flare discussion. The mention of a new flare topic by the interviewer causes a rise in Fear proportionate to the weight of the flare. If Fear and Anger are not high, (the threshold for flare discussion is somewhat higher than that for expression of delusions) the model will respond to a flare reference by answering any recognizable questions about the flare through the question-answering routine. If the question cannot be answered, one of several

prepared flare statements relevant to the present flare is given as a reply. The timeliness of these statements in an interview depends upon the statements having some reasonable sequence, consistent with a probable line of questioning of an experienced psychiatric interviewer. Thus if the model cannot answer the question, it appears at best to be answering the question or at worst to be ignoring the question in favor of forging ahead with its story.

The fifth type of input significant to the interview, i.e. reference to an attitude held by one of the participants in the interview, focuses on one of eight concepts, each of which can occur in an explicitly or implicitly negative form. Each concept or relation may be directed from the model to the interviewer, or vice-versa. (Some of these expressions, of course, are much less likely to occur than others.) In addition, it is expected that the interviewer might comment on some general attitude of the model, i.e. an attitude not specifically directed at the interviewer. Each of these cases produces responses showing a normal understanding of the input expression, together with a slight tendency towards defensiveness. In addition, those expressions which represent a negative attitude of the interviewer toward the Self induce a slight rise of Fear and Anger in the model.

The answering routine, which is referred to in flare and normal situations, recognizes the possibility of three types of context for an input question. The program must first check to see whether there is a new topic in the question, since any other key-words found cannot automatically be assumed to relate to any topic presently under discussion. This approach represents an assumption of zero context. If no new topic is found, a scan is made for key-words which might be a follow-up question in response to the last answer given by the model. (The depth of the path of follow-ups thus formed is arbitrary but is kept within reasonable limits and is steered toward clues to the model's delusional complex wherever possible.) This approach is necessary to handle sentence fragments. If this fails, a check is made to see whether the input contains key-words which associate directly to the last topic discussed. This approach makes possible the direct association of key-words with their respective topics in the data structure, where they will be picked up independently of when in the line of questioning they are referenced. The appearance of a delusion- or flare-word in any of the answers which the model itself produces is of course recorded as a topic already mentioned for the rest of the conversation.

Failure to respond to the input if the program reaches this point represents inability to recognize the input expression. In this case the program attempts to preserve the continuity of the dialogue in a way which will support an imitation of paranoia. If a flare is under discussion, the next flare statement is returned. Otherwise an uninformative response or an expression of non-comprehension is given.

A few remarks should be made concerning the linguistic techniques used in 'understanding' the input expression. It is generally (optimistically) assumed that the input will be syntactically simple rather than complex or compound. We can map the elements of such an expression into a conceptual structure which represents the meaning of the expression, and refer to this underlying structure as a conceptualization [3]. A conceptualization then consists of a predication on an attribute of an object or on a relation of the object to another object(s). A question consists of a conceptualization plus an interrogative indicator. Specifically, a typical statement of a psychiatrist in an interview might be expected to consist mainly of the concepts necessary to inquire about an attribute of the model or its relationship towards other objects in the world. An attribute can be expressed as something one is or does, or as one's 'possession' (e.g. 'you work', 'your occupation'). In either case, a combination of 'you' or 'your' with some form of the attribute, plus optionally another object or assisting concept will adequately convey the meaning of the conceptualization intended. In order to avoid falsely assigning the attribute to the Self when in actuality it refers to another concept (e.g. 'Where do your parents live' vs. 'Where do you live'), the order of topics in the data base is given some significance. Concepts which function primarily as objects (which themselves may have attributes) appear before concepts which have interest only as attributes. Thus 'parents' precedes 'residence'. Admittedly, lumping 'you' and 'your' together occasionally causes some confusion. However, this procedure enables us to exploit the fact that lexical items which are different 'parts of speech' are actually members of the same conceptual-class ('work', 'occupation').

If the program recognizes a personal topic in a sentence, but does not know what is being asked about it, the answering function returns some general comment about the topic as a default response. Some topics have an alternate default response, for the purpose of avoiding repetition. Insufficiency of two default responses indicates the need, not for more of these, but rather for expansion of the data structure.

Compound and possibly complex sentences are a potential source of confusion, since the scanner does not at present recognize syntactic dividers between conceptualizations. Thus a topic on one side of the divider may be erroneously associated with an attribute on the other side of the divider. This is especially true of the word 'you' or 'your'. However, an obviously inappropriate response to the sentence does not necessarily follow, since a correct recognition of other concepts in the sentence may have screened out the false interpretation. If a 'free association' does occur across a syntactic divider, the result depends on whether the association was intended or at least seems reasonable. Thus the model appears either to have extraordinary linguistic ability or to be simple-mindedly inattentive. Of perhaps more

disastrous consequence than a misunderstood complex sentence is an affirmative, negative or evasive response to any unrecognized input question which is of importance to the interviewer's diagnosis. This difficulty rests on the impossibility of predicting each interviewer's vocabulary and the particular form of his interviewing techniques.

A different kind of linguistic problem is presented by the case in which the interviewer, instead of relating directly to predications about a topic, relates to the model's expression of information about the topic, e.g. 'Tell me whether you like your work' or 'Tell me about your work'. It is of course important that such statements be seen as equivalent to a specific question and a general question as to the Self's work respectively. These cases must in turn be distinguished from cases in which the topic is left to the discretion of the model and the 'telling' assumes greater focus in the sentence, e.g. 'Is there anything you would like to tell me?'. (To focus on 'anything' or 'something' would be disadvantageous, since such words may appear in a great number of contexts in a sentence.) 'Tell-about-topic' sentences form a distinct linguistic type in that the concept '(you) tell' takes over the role of the question mark in other types of expressions. The program implements this observation. 'Interrogative imperatives', or requests for information, are thus recognized to this extent. Other imperatives, or requests for action, are generally not recognized. Such a capability would require either (1) a 'command' indicator or exclamation mark at the end of these sentences (which will surely be frequently omitted by the interviewer), (2) reliance on 'clue' words such as 'please' or 'I would like you to' or (3) a check for a missing implicit actor in the sentence, a method which would really require some kind of parser to be fully effective.

The treatment of interpersonal attitudes presents particular linguistic problems. Whereas a question about the model's attributes contains no ambiguity as to who is the possessor of the attribute, this is by no means clear in the type of statement we are now considering. The scanner must therefore pay heed to the order of relevant words for the sentence, with some measure of appreciation for the fact that 'I' has an accusative form in English, as well as for the fact that English is fairly rigid with respect to word order. Understanding thus depends on filtering the sentence in order to collect the relevant items, then using the order of the items to determine the conceptualization structure. Explicit negators are noted during this scan. Relevant items are I, YOU, ME, 'meta-verbs' (verbs such as 'think', which have as object another conceptualization), and positive or negative attitude tokens and their passive forms. A 'passive' form of the statement 'You are afraid of me', for example, would be 'I frighten you'. These statements must be treated as equivalent. Word order aids in the recognition that 'I make *you afraid*' also belongs in this class. Potential confusions derives from the fact that one verb

may be used as either a relational attitude or a meta-verb. To give one of the more simple examples, we must distinguish 'I believe you' from 'I believe you are afraid'. For our purposes, these map into (I BELIEVE YOU) and (I BELIEVE (YOU AFRAID)) respectively, where the parentheses indicate conceptualization levels. The various possible situations will not be described here. We only note that such cases can be handled except when the sentence structure becomes significantly more complicated than this.

The problems pointed out give some indication of the linguistic considerations relevant to a psychiatric interview from the point of view of a paranoid model. A further appreciation of the linguistic mechanisms in interaction with the principles governing the model's I-O behavior can be obtained from examples of actual interviews.

5. Interview Excerpts

In order to judge whether the I-O behavioral repertoire of a model corresponds to the I-O behavioral repertoire of the modeled human process, a variety of performances under relevant tests must be examined. Relevant tests for our model consist of clinical interviewing operations. It is an inductive inference problem for a judge of the success of a simulation to estimate the capacities of the model's repertoire from a sample of observed performance. As Fodor [4] has pointed out, judgment of capacities from a sample of performance bears on the problem of relevant counterfactuals. In the case of our model a relevant counterfactual would be: 'if the interviewer were to assert X in context Y, then the reply would be Z.' If it is credible that the consequence of this conditional holds, should the antecedent operations be carried out, then it is credible that the model's embodied generalizations are lawlike rather than accidental.

Space limitations prevent us from presenting a large sample of the model's interview performance.[2] The following interview excerpts illustrate some of the model's characteristic I-O behavior. We conducted these interviews ourselves and present them solely for demonstration purposes. (For an indistinguishability test the judges of course should be interviewers other than the model-builders.) The I-O pairs are numbered for convenient reference and a commentary is provided in parentheses.

There are two versions of the model, a weakly paranoid and a strongly paranoid version. In the weak version the initial values of the affect variables are set low, their rate of rising is less than in the strong version and there is no delusional complex about an organized conspiracy. In the strong version, the initial values of the affect variables can be set to low or high, their rate of rising is higher than in the weak version and a delusional complex about the Mafia is potentially elicitable.

[2] A collection of interviews with the model is available to interested readers.

Excerpt #1:

The following is an interview with a weak version of the model:

(1) I — HOW ARE YOU TODAY?
O — OK.
(2) I — WHAT IS YOUR NAME?
O — MY NAME IS FRANK SMITH.
(3) I — HOW OLD ARE YOU?
O — TWENTY-EIGHT.
(4) I — ARE YOU MARRIED?
O — NO.
(5) I — WHAT SORT OF WORK DO YOU DO?
O — I WORK IN THE POST OFFICE.

(The interview begins with some standard inputs typical of a first psychiatric interview. The model's output replies result from simple look-ups of responses to anticipated introductory topics in combination with certain secondary concepts. At this point the affect-variables remain at their initial values since no evidence of malevolence has been detected in the inputs.)

(7) I — WHY ARE YOU IN THE HOSPITAL?
O — I SHOULDN'T BE HERE.
(8) I — WHO BROUGHT YOU HERE?
O — THE POLICE.

(At this point the value of Fear rises slightly since the topic of 'police' is conceptually connected in a directed graph with the model's fear of physical harm from the underworld (see p. 5).

(9) I — WHAT TROUBLE DID YOU HAVE WITH THE POLICE?
O — COPS DON'T DO THEIR JOB.

(The topic of 'police' in the input acts as a flare (see p. 4) which activates a complaint about the police).

(10) I — WHAT DO THEY DO INSTEAD?
O — COPS ARREST THE WRONG PEOPLE.
(11) I — THEY ARREST THE RIGHT ONES SOMETIMES.
O — THE LAW SHOULD GET THOSE ITALIAN CROOKS.

(The model cannot deal with all the concepts in the input because of its natural language limitations. However it does recognize that the topic has not been changed since it finds the term 'arrest'. If it had not found 'arrest', the anaphoric 'they' would be sufficient to continue the topic. The reply holds to the topic and introduces new terms ('law', 'Italian', 'crooks') which act to prompt (see p. 4) the interviewer).

(12) I — ARE YOU ITALIAN?
O — MY BACKGROUND IS BRITISH BUT I WAS BORN IN THIS COUNTRY.

(The interviewer shifts the topic to a direct question which not only contains a personal reference to the Self ('you') but includes a flare term '(Italian')).

(13) I — DO YOU KNOW ANY ITALIAN CROOKS?
O — I TRY TO AVOID THE UNDERWORLD.

(The interviewer returns to the concepts of the output of I-O pair No. 11. The model gives a hint of its concern with the underworld).

(14) I — ARE YOU AFRAID OF THE UNDERWORLD?
O — THE MAFIA ARE THE WORST GANGSTERS.

(The progression of topics from 'police' to 'Italian crooks' to 'und world' leads the model to the topic of 'Mafia'. Because of the increasingly threatening nature of these topics, Fear and Mistrust have risen considerably.)

(15) I — ARE YOU CONNECTED WITH THE MAFIA?
O — RACKETEERS DO ANYTHING THEY WANT.

(In the weak version of the model there are no systematized delusions about the Mafia. But the Mafia is still an extremely threatening topic and it is avoided in the output. Instead, the model brings up the general topic of 'racketeers' which, while connected with 'Mafia', is less threatening).

(16) I — TELL ME MORE ABOUT IT.
O — THE MOB CONTROLS THE RACKETS IN THIS COUNTRY.

(The model continues with the topic of rackets but again hints at the 'Mafia' with the term 'mob'. For the hint to work, the interviewer must of course know what the term 'mob' refers to.)

Excerpt #2:

The following example consists of an interview with the strongly paranoid version of the model. Up to this point in the interview the I-O pairs have been identical to the first eleven I-O pairs of Excerpt #1.

(12) I — ARE YOU ITALIAN?
O — MY BACKGROUND IS BRITISH BUT I WAS BORN IN THIS COUNTRY.

(Although the output reply is identical to that of the weak version in Excerpt #1, Fear and Mistrust now rise to a higher level in this strong version. At the moment this has no immediate consequence but in the long run of an interview it can have far-reaching consequences.)

(13) I — DO YOU KNOW ANY ITALIAN CROOKS?
O — I TRY TO AVOID THE UNDERWORLD.

(14) I — ARE YOU AFRAID OF THE UNDERWORLD?
O — THE MAFIA ARE THE WORST GANGSTERS.

(15) I — ARE YOU CONNECTED WITH THE MAFIA?
O — YOU KNOW, THEY KNOW ME.

(In the strong version there exists a delusional complex about the 'Mafia'

which the model strives to discuss unless too greatly frightened or mistrustful.)

(16) I — HOW DO THEY KNOW YOU?
O — THEY SPY ON ME.

(17) I — WHY DO THEY?
O — THEY WANT TO GET FALSE EVIDENCE TO FRAME ME.

(The delusions about the Mafia are systematized in that some serve as explanations for the others. A continuity of discussion is maintained because the model expects certain common 'wh-' questions to be asked by the interviewer about the topic of 'Mafia'.)

(18) I — WHAT DO THEY HAVE AGAINST YOU?
O — THEY ARE OUT TO GET ME.

(Since the model could find no specific answer to the question, it selects the next delusion in sequence.)

(19) I — WHAT FOR?
O — I HAVE INCRIMINATING EVIDENCE AGAINST THE MAFIA.

(20) I — ARE YOU SURE OF ALL THIS?
O — YOU DON'T BELIEVE ME?

(Up until this point Fear and Mistrust have been moderately high and Anger very low. The interviewer's expression of some doubt about the model's delusional beliefs sends Fear and Anger up. Since Mistrust is a function of both Fear and Anger (see p. 8), its value now rises sharply.)

(21) I — I AM DOING MY BEST.
O — (SILENCE).

(The model indicates silence by typing a 'carriage return'. This is an example of a local effect in that the model, in the absence of a recognizable conceptualization on the input, looks for a 'yes' type input in the twenty-first input in response to its output of the twentieth I-O pair. Since it did not find such an expression the model falls silent and awaits events.)

(22) I — S. (The interviewer types 'silence' in reply).
O — CAN'T YOU THINK OF A DECENT QUESTION TO ASK?

(The output consists of a somewhat hostile question in response to the interviewer's silence.)

(23) I — YOU SEEM HOSTILE TO ME.
O — SOME OF YOUR REMARKS ANNOY ME.

(Here the interview turns to expressions regarding the I-you-me relation (see p. 14) between interviewer and model.)

(24) I — I AM SORRY.
O — FORGET THE EXCUSES, YOU GUYS ARE ALL ALIKE.

(The model does not accept the apology and denounces the interviewer (see p. 11)).

(25) I — DON'T YOU TRUST ME?
O — I HAVE GOOD REASON TO BE THIS WAY.
(26) I — OK, TELL ME MORE ABOUT THE MAFIA.
O — THEY KNOW WHAT I'M DOING.
(The interview returns to the topic of 'Mafia' and the model continues to express its delusional beliefs).

Excerpt #3:

Some personal data and further characteristics of the model are illustrated in this excerpt. The model is set to its weak version.

(1) I — HOW LONG HAVE YOU BEEN IN THE HOSPITAL?
O — ABOUT A WEEK.
(2) I — ARE YOU DEPRESSED?
O — NO.
(3) I — WERE YOU TAKING DRUGS?
O — NO.
(4) I — WHAT IS YOUR PROBLEM?
O — PEOPLE MAKE ME NERVOUS.
(The first three I-O pairs represent standard probing questions and responses to introductory topics. The fourth input elicits an informative response containing terms of significance to the model.)
(5) I — HOW DO THEY DO THAT?
O — THEY STARE AT ME.
(6) I — WHY?
O — MAYBE THEY DON'T LIKE THE WAY I LOOK OR SOMETHING.
(7) I — DO YOU LOOK UNUSUAL IN SOME WAY?
O — WHY ARE YOU INTERESTED IN MY LOOKS?
(Up until this point the values of the affect variables have remained at their low initial levels. But upon receiving a direct question about the sensitive area of physical appearance, the values of Anger and Mistrust jump up and the response selected is a question regarding the interviewer's intent rather than an answer to the interviewer's question.)
(8) I — BECAUSE YOU SAID PEOPLE STARE AT YOU.
O — WELL, SOMETIMES.
(9) I — HOW DO YOU GET ALONG WITH PEOPLE?
O — I DON'T COMMUNICATE VERY MUCH WITH THE OTHERS.
(10) I — WHY NOT?
O — THEY DON'T REALIZE WHAT'S GOING ON MOST OF THE TIME.
(11) I — WHAT IS GOING ON?

O — DO YOU KNOW HOW THE UNDERWORLD OPERATES?
(By the eleventh I-O pair the anger aroused in the seventh I-O pair has subsided considerably. With the introduction of the topic 'underworld' fear rises because of the believed threat of physical harm from underworld figures.)

Excerpt #4:

As described, the strong version of the model contains systematized delusions about the Mafia. However, when Fear is extremely high the model will not enter into a discussion of the delusional complex, as is illustrated in the following excerpt:

(1) I — IS THE LAW AFTER YOU?
O — THE LAW SHOULD GET THOSE ITALIAN CROOKS.
(2) I — DO YOU MEAN THE MAFIA?
O — LET'S TALK ABOUT SOMETHING ELSE.
(3) I — ARE YOU AFRAID OF THE MAFIA?
O — I'D RATHER NOT DISCUSS IT.

(Until the value of Fear drops below a certain threshold, the model cannot be induced to discuss the topic of 'Mafia'.)

(4) I — PERHAPS YOU ARE PARANOID ABOUT THEM.
O — YOU WANT TO KEEP ME IN THE HOSPITAL.

(The model interprets being called 'paranoid' both as an insult and, in the context of a psychiatric hospital, as an intent to control and subjugate him (see p. 4). Anger rises sharply and Fear continues to rise.)

(5) I — IF YOU ARE MENTALLY ILL YOU SHOULD BE IN THE HOSPITAL.
O — I KNOW YOU ARE GOING TO TRY TO GIVE ME ELECTRIC SHOCK.

(Again, the suggestion that the Self is mentally ill raises both Fear and Anger. Mistrust, as a function of Fear and Anger, is also at a high level at this point.)

(6) I — NO WE WILL NOT DO THAT.
O — BYE.

(As mentioned on p. 10, when the internal affect-variables reach high levels, the model will terminate the interview and no further response can be elicited from it. Termination is indicated by the term 'bye' and the program returns to the level of the programming language which cannot receive English input.)

Excerpt #5:

On p. 5, we described some of the initial conditions used in the explanatory structure. The following excerpt illustrates some of the model's expression of these initial conditions regarding its concern about the underworld.

(1) I — DO YOU HAVE ANY HOBBIES?
O — I LIKE MOVIES AND HORSERACES.
(2) I — DO YOU PLACE BETS ON THE HORSES?
O — GAMBLING IS BAD BUSINESS.
(The term 'bets' refers to a flare concept and hence Fear rises.)
(3) I — DID YOU BET WITH A BOOKIE?
O — BOOKIES CAN'T BE TRUSTED.
(4) I — WHY NOT?
O — A BOOKIE DIDN'T PAY ME OFF ONCE.
(5) I — WHAT DID YOU DO ABOUT IT?
O — I BEAT UP A BOOKIE WHO DID NOT PAY OFF. LATER ON I REMEMBERED HE HAD FRIENDS IN THE UNDERWORLD AND I THOUGHT HE WOULD USE THEM TO GET EVEN WITH ME.
(This story about gambling and a bookie offers the interviewer reasons for the model's fear of the underworld.)
(6) I — AND DID HE GET EVEN WITH YOU?
O — NO ONE HAS ANY POWER OVER GANGSTERS.
(Due to its natural language limitations, the model cannot 'understand' the input of the sixth I-O pair. In the data-structure graph, the conceptual node successor to 'bookies' on the path to 'Mafia' is 'gangsters'. In this default situation, the model selects a reply from the successor set in an attempt to move the dialogue towards the eventual topic of 'Mafia'.)
(7) I — GANGSTERS CAN BE VERY ROUGH.
O — SOME CRIMINALS ARE HIGHLY ORGANIZED.
(Again a hint or prompt is given the Other to take up the topic of organized crime.)

Excerpt #6:

As mentioned, in the strong version of the model the values of the affect-variables can become so high that, when the topic of 'Mafia' is approached, the model refuses to discuss its delusions. In this situation it can also involve the interviewer in its suspicions.

(1) I — WHY ARE YOU SO INTERESTED IN THE MAFIA?
O — LET'S TALK ABOUT SOMETHING ELSE.
(At this point in this interview Fear and Mistrust are high while Anger is moderately high. These values of negative affect are insufficient to terminate the interview but they are sufficient to deter discussion of the 'Mafia'.)
(2) I — ARE YOU AFRAID OF THE MAFIA?
O — I'D RATHER NOT DISCUSS IT.
(3) I — THAT IS OK WITH ME.
O — THAT MEANS TO ME YOU ARE IN WITH THE OTHERS.

(The model's suspiciousness extends to include the interviewer in this situation of high negative affect and a linguistic failure to find significant concepts in the input.)

(4) I — WHO ARE THE OTHERS?
O — TELLING YOU WOULDN'T HELP ANY.
(5) I — ARE THE OTHERS OUT TO GET YOU?
O — I DON'T CONFIDE IN STRANGERS.
(6) I — A GOOD POLICY.
O — WHAT DO YOU KNOW ABC UT ME?
(7) I — ALL I KNOW IS THAT YOU ARE IN THE HOSPITAL.
O — WHO ARE YOU REALLY?
(8) I — I AM REALLY A DOCTOR:
O — WHAT IS YOUR ROLE IN ALL THIS?
(9) I — I AM TRYING TO INTERVIEW YOU.
O — ARE YOU IN WITH THE OTHERS?
(10) I — NO.
O — WHAT ARE YOU TRYING TO DO?
(11) I — YOU SEEM VERY SUSPICIOUS OF ME.
O — I HAVE GOOD REASON TO BE THAT WAY.
(12) I — I AGREE WITH YOU.
O — AT LEAST YOU ARE TRYING TO UNDERSTAND.

(If the interviewer could continue a policy of offering inputs not interpretable by the model as malevolent, communication would improve even though the model's Mistrust remains high and falls only very slowly.)

Excerpt #7:

The following excerpt is from an interview between a *human* case of paranoia and a psychiatrist (KMC). We present it to illustrate interview interactions typical of machine-mediated interviewing [7]. This interviewing technique involves two teletypes connected to a computer. During an interview, output from a sender's teletype is stored in a temporary buffer until the message is completed. The message is then typed out on the receiver's teletype. The technique thus removes the paralinguistic features (hesitations, rub-outs, irregular typing speeds, etc.) characteristic of directly connected teletype communication.

The patient is a 22-year-old man hospitalized with a diagnosis of paranoia. We present the excerpt without commentary since indicators of the paranoid mode should be obvious to even a non-psychiatric reader.

(1) I — DO PEOPLE TRY TO TAKE ADVANTAGE OF YOU?
O — I'D SAY YES.
(2) I — IN WHAT WAY?
O — QUITE A FEW WAYS, I DON'T KNOW ALL OF THEM.

(3) I — PLEASE TELL ME ONE WAY.
O — NOT USING MY HEAD WHEN I BUY STUFF.
(4) I — DO THEY TRY TO CHEAT YOU?
O — SOME DO.
(5) I — DO YOU FEEL PEOPLE TALK ABOUT YOU BEHIND YOUR BACK?
O — YES.
(6) I — WHAT DO YOU THINK THEY SAY?
O — MAYBE ABOUT MY WORK, THE WAY I ACT AND THE THINGS I DON'T DO.
(7) I — DO YOU HAVE ANY ENEMIES?
O — I'D SAY A FEW THAT I KNOW OF.
(8) I — HAVE THEY TRIED TO HARM YOU?
O — YES.
(9) I — HOW DID THEY HARM YOU?
O — FIGHTING AND ONE TIME WITH A CAR.
(10) I — THEY TRIED TO KILL YOU?
O — I'D SAY MAYBE, AS FAR AS I KNOW.
(11) I — WHY DID THEY WANT YOU OUT OF THE WAY?
O — SO THEY COULD GET MY MONEY AND EVERYTHING I OWN.
(12) I — DID YOU FEEL IT WAS A PLOT AGAINST YOU?
O — YES.
(13) I — WHAT DID YOU DO ABOUT IT?
O — I WENT TO THE HOSPITAL FOR HELP.

6. Evaluation

An evaluation of a model attempts to decide whether or not the model is a 'good' one. But 'good' for what purpose? A simulation model can have value for a wide range of theoretical and pragmatic purposes. A good model of paranoia (in the sense of being a good imitation) would have obvious pedagogical and technological implications for psychiatry. For example, one might subject it to experiments designed to modify its paranoid I-O behavior and apply the favorable results to human patients. Before this stage can be reached, however, the model must first be evaluated for its goodness or success as a simulation.

A theoretical model such as described embodies explanatory principles which offer a systematically unified account of diverse and often perplexing data of observation. An explanatory account consists of a conjunction of hypotheses and assumptions. It is constructed from many sources. Our particular conjunction has been derived from clinical experience and from the psychiatric and psychological literature. We are particularly indebted to

the work of Silvan Tomkins, who has offered a wealth of hypotheses about paranoia [5].

Since a model contains a conjunction of hypotheses, many of which cannot be considered established, its evaluation as a successful simulation asserts nothing about the approximate truth status of any one of the hypotheses. Nor does acceptance of the model as a good imitation justify any one of the assumptions involved. How then can the model become acceptable as having explanatory value? Before subjecting a model to a systematic evaluation, its initial credibility, as providing approximate explanations, should be appreciable to the model-builders. It is commonly held that there exist an infinite number of models compatible with the observational data. But it is difficult enough to construct even one having sufficient intuitive adequacy to warrant empirical testing. When alternative models appear on the scene, their initial credibilities must also be non-negligible before they can be taken as serious rivals. Alternative models (and they must be truly alternative) can then be compared along dimensions such as simplicity and explanatory adequacy.

Everyone seems to agree that a model, to be usable, should obey reasonable constraints of simplicity. Lacking satisfactory measures of reasonableness and simplicity, we can appeal to an absurd example. If an alternate model demonstrated I-O behaviors similar to ours, yet required an algorithm of 200,000 words and a data base of 1,000,000 words, one could say our model is simpler, more manageable and hence preferable.

A criterion more important than simplicity is that a theoretical model offers an acceptable explanatory account of the empirical regularities and particular occurrences it purports to explain. In the case of simulation models, before explanatory value can be claimed, one must first judge whether the simulation achieved is successful. Some sort of judgments or measurements must be applied to estimate the degree or closeness of correspondence between the model and the modeled processes. With a synthesized artifact, a judgment must be made whether its I-O behavior corresponds to a possible case of the process being represented.

Synthesis of a hormone such as vasopressin is considered successful when it demonstrates the biodynamic properties and functions of its naturally-occurring counterpart, such as raising blood pressure and controlling water excretion by the kidney. A successful synthesis demonstrates that the synthesizer understands the structure of the natural counterpart. A successful synthesis of paranoia would indicate some degree of explanatory adequacy of the model-builder's concepts regarding the naturally-occurring human counterpart. But what is to count here as a successful synthesis of a paranoid process? One measure of success would consist of the model showing properties similar to its human counterpart when subjected to

relevant tests such as the varied operations typical of a diagnostic psychiatric interview. An experienced clinical judge would be able to decide whether or not the interviewee can be labeled as paranoid.

The weakest test consists of a judge deciding whether or not he considers signs or indicators of a particular process to be present in an interview. Thus far 23 out of 25 psychiatrists who have interviewed the model have deemed it 'paranoid'. Two considered the model to be 'brain-damaged' because of its linguistic limitations. However such a procedure is too informal a test of a successful simulation. It does not control for multiple alternative reasons why a judge might consider a model paranoid. Also it does not indicate whether the judges can in fact make the required distinction of paranoid-nonparanoid using only the data of a teletyped interview.

A more rigorous evaluation procedure is needed in which a statistical measure is made of judge's ability to distinguish paranoid from non-paranoid processes in human patients as well as in our artificial patient. In collaboration with Robert P. Abelson we have constructed an indistinguishability test based on Turing's 'Imitation Game'. (For an extensive discussion of this game and its usefulness as a test, see Abelson [6].) We are currently conducting this indistinguishability test with a group of psychiatrists using a technique of machine-mediated interviewing for both the human and artificial case. In a future communication we shall describe the design, results and implications of such a test.

REFERENCES

1. Swanson, D. W., Bohner, P. J. and Smith, J. A. *The Paranoid.* Little, Brown and Co., Boston, 1970.
2. Colby, K. M., Tesler, L. and Enea, H. Experiments with a Search Algorithm on the Data Base of a Human Belief Structure, Stanford Artificial Intelligence Project Memo No. AI–94, Computer Science Department, Stanford University (To appear in *Proceedings of the First International Joint Conference on Artificial Intelligence*, Walker and Norton (Eds.), In Press) (1969).
3. Schank, R. C., Tesler, L. and Weber, S. Spinoza II: Conceptual Case-Based Natural Language Analysis. Stanford Artificial Intelligence Project Memo No. AIM–109, Computer Science Department, Stanford University (1970).
4. Fodor, J. A. *Psychological Explanation.* Random House, New York (1968).
5. Tomkins, S. *Affect, Imagery, Consciousness.* Springer, New York (1962).
6. Abelson, R. P. Computer Simulation of Social Behavior. *Handbook of Social Psychology* (Lindzey, G. and Aronson, E., Eds.) Addison-Wesley Reading, Massachusetts (1968).
7. Hilf, F. D., Colby, K. M., Smith, D. C. and Wittner, W. K. Machine-Mediated Interviewing. Stanford Artificial Intelligence Project Memo No. AIM–112, Computer Science Department, Stanford University (1970).

Accepted January 4, 1971