

# Modeling a paranoid mind

**Kenneth Mark Colby**

Neuropsychiatric Institute, School of Medicine, University of California,  
Los Angeles, Calif. 90024

**Abstract:** Algorithmic model building in artificial intelligence provides a contemporary method for explaining certain kinds of human behavior. The paranoid mode of thought and action represents a kind of pathological behavior that has been well recognized for over twenty centuries. This article describes a computer simulation model embodying a theory that attempts to explain the paranoid mode of behavior in terms of strategies for minimizing and forestalling shame-induced distress. The model consists of two parts, a parsing module and an interpretation-action module. To bring the model into contact with the conditions of a psychiatric diagnostic interview, the parsing module attempts to understand the interview input of clinicians communicating in unrestricted natural language. The meaning of the input is passed to an interpretation-action module made up of data structures and production rules that size up the current state of the interview and decide which (linguistic) actions to perform in order to fulfill the model's intentions. This module consists of an object system which deals with interview situations and a metasystem which evaluates how well the object system is performing to attain its ends. The fidelity of the simulation has been tested using Turing-like indistinguishability tests in which clinical judges attempt to distinguish the paranoid model-patient from an actual paranoid patient. Since clinicians are unable to make the relevant distinctions, the simulation is considered successful at the input-output level of functional equivalence. Issues of underlying structural equivalence and the nature of generative pattern explanations are discussed in the light of the model's potential value in guiding clinical decisions and intervention strategies in paranoid disorders.

**Keywords:** Affect; artificial intelligence; computer simulation; cognition; language; mind; paranoia; psychopathology; reasoning

## 1. Introduction

In this article I propose to describe an area of artificial intelligence (AI) research that I and several colleagues have been engaged in for a number of years. This work concerns the problems of designing, constructing, and testing a model of paranoid thinking.

Beginning with some general issues about the cognitive sciences and artificial intelligence, I then proceed to the specific problem of paranoia and a description of a theoretical simulation. Evaluation of the model and a few more general issues are taken up at the end.

### 1.1 The cognitive sciences

With the growth of the contemporary consortium of cognitive sciences (science itself is a cognitive art), including psychology, linguistics, artificial intelligence, philosophy, and the neural sciences, the concept of mind is enjoying renewed popularity after being outlawed by behaviorism for several generations (Colby 1978; Fodor 1975; Simon 1980). Traditionally, psychology has been defined as the study of human behavior. But the behavior of what? If the whole person is taken as the agent, then descriptions are in person language using persons' names or pronouns as subjects of propositions (as in "Tom cannot remember his uncle's name"). If the person is partitioned into psychological components and processes (memory, belief, affect, and so on), the person language changes to an impersonal language with a change of subject, as

in, "Tom's memory is failing. *It* cannot retrieve his uncle's name." Our descriptive vocabulary may still contain proper names as modifiers but the explanatory vocabulary now involves the impersonal qualities of an it-language. This shift of subject means that the observable behavior to be explained must be accounted for by the unobservable behavior of a structure of components and processes. Observable behaviors are perceptual patterns for which hypotheses are invented and by which hypotheses are tested. They initiate and terminate cycles of inquiry regarding the central entities of interest, the transobservational generating (nonmechanical) mechanisms that produce the observable behavior patterns in question.

### 1.2 Taxonomic and explanatory knowledge

The first step in science is to know one thing from another (Linnaeus). We want to know what kinds of entities there are in the world and what they tend to do. Kinds (classes, groups, taxa, types) eventually serve as bound variables in laws. Under "mentalese" (Fodor 1975) descriptions of "person," the kinds that have been of interest in psychology and philosophy have been termed "wants," "beliefs," "emotions," "purposes," and the like. The vocabulary of this taxonomy remains close to that of everyday folk psychology. However ramshackle, this unsystematic folk psychology does make sense of, and correctly anticipates, a great range of human behavior (Churchland 1979; Pylyshyn 1980). Thousands of agreements are reached

in everyday life about both the meaning and the truth of statements using mentalistic terms. Why cannot a scientific psychology, with a privileged vocabulary, develop some measure of consensus from such a roughly dependable, although superficial and error-prone, starting point?

What are we trying to explain? A broad description comes from Harre (1977, p. 155). "The given phenomena requiring to be understood, are plan-forming and rule-following by active, self-monitoring agents, the understanding of meaning, the performance of actions meaningful to others, and the mutual production of patterned, integrated action-sequences in the performance of which social acts are achieved." This is indeed a tall order.

Scientific explanations are achieved through the use of theories and models that have been extensively treated in a long literature (Leatherdale 1974). A theory is a hypothetico-deductive system whose hypotheses connect and cohere with one another in a noncontradictory way. Each hypothesis stands as an initial assumption or as a consequence of one or more initial assumptions. For a theory to be testable, it must be empirically sensitive to observations that can strengthen or weaken its acceptability. Many mentalistic theories are formulated in natural language, whose vaguenesses and ambiguities are well known. Such formulations, as a large set of statements, can be difficult to analyze because we cannot tell if the assumptions are independent or redundant; we cannot tell if each assumption is needed, or whether the assumption set is complete or incomplete; we cannot easily decide on the logical status of the assumptions — are they axioms, tautologies, definitions, typologies, or empirical statements?

In theoretical models, simplifications and idealizations are introduced. Simplification involves selecting only centrally relevant processes. Idealization involves assuming theoretically exact classes (in contrast to inexact empirical classes) and perfect properties. These simplifications and idealizations introduce errors of omission and commission in theoretical models.

In what way, then, does an algorithmic model explain? An algorithmic model's explanatory power is related more to its generative power than to its predictive power. Such an explanation unifies observable input/output patterns by hypothesizing an internal structure of effective computational patterns connecting themselves and the input/output patterns into a comprehensive and organized whole. Such an explanation attempts to account for the existence, persistence, and change of observable patterns by showing how they belong to, or are the product of, a finite structure of higher-level, transobservational patterns that encompass a multiplicity of patterns at the observational level. Thus we can understand an entity not just through a list of its local properties but by identifying it as a specific part of an organized whole, consisting of patterned relations and structures.

### 1.3 Artificial intelligence

As one of the newer cognitive sciences, artificial intelligence (AI) provides novel concepts, methods, and

instruments for theory construction through algorithmic model building. Let us consider a few special issues concerning AI's theories and models.

As a type of theoretical psychology, AI takes as its entities of inquiry the representations of mental processes. Its vocabulary is a mixture of terms from ordinary folk psychology ("belief"), scientific psychology ("long-term memory"), linguistics ("phrase-structure grammar"), and computer science ("LISP atoms"). Its problems concern the computational mechanisms effective in generating observable behavior. It has methodological do's ("do design, write, and run algorithms"; "do construct theories and experiments with algorithms") and don't's ("don't collect any more data because we already have more than we can understand"). Its models are algorithmic, consisting of (a) a programming language in which patterns of rules and representations can be rigorously specified, (b) an algorithm composed of procedures, data structures, and supplementary implementations needed to get the algorithm to run on (c) a machine processor that can rapidly and reliably carry out the computations specified by the algorithm. Successful runs of the algorithm provide tests of consistency. Comparing the input/output behavior of the algorithm with that of people in the relevant reference class being modeled may provide one empirical test of adequacy. AI theories are stuff-free, detached from any particular physical realization, but constrained by the concept of a virtual machine.

In AI, in my view, it should be possible to state a theory independently of its model, so that those who do not understand the programs can understand them in a nonprogramming language. Theories and models are two ways of manifesting our theoretical concepts. In the case of computational models, there are three further distinctions I draw between theory and model. (1) All the statements of a theory are intended to be taken as true, or as candidates for truth, whereas some statements in the programming language in the model, features necessary to get the model to run, are intended to be read as false, or at least of uncertain truth value. Thus, a computer-theoretic model, or, more precisely, a listing of the program, in order to be properly read and publicly scrutinized, must be accompanied by a commentary provided by the investigator so that implementation details, such as multiple right parentheses, are distinguishable from the principles of the theory embodied in central constructs of the model. For example, in the model for paranoia described in section 2, numbers are used to represent values of affect variables. The numbers are a feature of the *model*, involved in its execution, but they do not represent quantitative values asserted by the *theory* of paranoia being realized. (2) A theory states that its entities have a certain structure, but unlike an algorithmic model, it does not exhibit that structure in itself. A model instantiates a kind, whereas theory statements use noninstantiative symbols to describe a kind. (3) A theory is read but not run; in other words, it remains in its initial state as a set of statements, whereas an algorithm, realizing a theory (which interfaces with a person, say through natural language dialogue), is dynamic in that it changes its states over time, some

states affecting other states, terminating in a final total state different from its initial total state.

Conventional psychology starts "bottom up" with simple units and tries to build them into systems that can account for complex behavior. (George Miller has quipped that the psychological literature looks like a small-parts catalog for a machine that has not yet been built.) This traditional approach has not been much of an improvement over folk psychology so far. In contrast to a conservative psychology, the method of AI constitutes a bold attempt at a constructive and compositional "top-down" strategy which starts with the question, What does a system have to have in order to produce the action patterns of type X, Y, Z? This approach is transcendental in Kant's sense, postulating what is necessary for some kind to be possible. For example, if a system is of a kind that can carry on a natural language dialogue, it must have a lexicon of words connected to concepts with meaning. Starting with higher-level "chunked" units, AI workers design a model and then run the algorithm to see how it works. Bottom-uppers wonder how one can even dare to model all sorts of complex mental processes when we do not understand the simplest of functions, such as short-term memory. As the famous engine designer said, however, "Start her up and see why she don't go!" Running the model represents a controlled Gedankenexperiment put to nature. But it has to be put to nature in a language that nature understands and that allows an unambiguous reply. Does the model do what it is supposed to do? Does it make sense? Does it reproduce the patterns of observation being investigated? Does it achieve the desired effects, as judged by our experience with the kind of entity being modeled?

"Theory construction is primarily model building, in particular imagining paramorphs" (Harre 1970, p. 46). To understand this quotation, let us first try to clarify what is meant here by analogy, simulation, representation, model, and their associated concepts (following Bunge, 1973, to a large extent).

Entities (a primitive concept too important and indispensable to be defined) are factual ("things") or conceptual ("numbers"). When two entities, X and Y, share properties, they are taken to be substantively analogous. If there exists a formal correspondence between properties of X and Y, we take them to be formally analogous. (Isomorphism is perfect formal analogy found only in mathematics and not in the factual sciences.) The analogy (similarity) relation is symmetric (X is analogous to Y and Y is analogous to X) and reflexive (X and Y are analogous to themselves). But the analogy relation is not necessarily transitive or intransitive. When the similarity of an analogy is propagated, the relation is transitive; that is, X is similar to Y which is similar to Z. X, Y, and Z then form an equivalence class, which is stronger than simple analogy.

To simulate an entity is to imitate it in only some of its aspects, that is, its kinds. The simulation relation is binary, symmetrical, reflexive, and transitive. Thus, it is an equivalence class as defined above. When a computer program simulates human behavior, it imitates aspects of (not all of) that behavior under the particular description of a kind. The imitation is

achieved through a representational relation, in other words, a model. A representational relation need not be a symmetric similarity relation but can be a symbolic proxy relation in which X does not resemble Y but stands in for Y for some agent Z, who uses X for purposes P. Confusion arises when Y is itself a representation, a conceptual entity or model taken by an investigator as a factual entity when constructing a model of a kind of human model representation.

Some models are empirically tested by pairwise comparison of the observable input/output behavior of the model with that of the kind of entities being modeled. Conceptual, rule-governed, internal representational behavior cannot be directly observed. Its properties must be detected or inferred from observations. In the paranoid model described below, the input/output simulation takes place in the form of natural language dialogue. To test the adequacy of the model, expert judges compare their experience with the linguistic responses of the model to their experience with responses from interviewed patients characterized as members of the kind "paranoid." If the interview similarities of input/output action types are sufficient, test judges, using their recognitional capacities (1) diagnose the model as paranoid and (2) cannot, under relevant constraints, distinguish the model from real patients (Colby 1975; Heiser, Colby, Faught & Parkinson 1980). One could construct a model of paranoid thought that did not use natural language dialogue. Ways other than interviewing would then be required to confront the model with perceptual, recognitional experience.

Construction of a performance model of kinds of behavior typical of a reference class presents certain difficulties for outside observers who wish to evaluate the model's merits and deficiencies. In a model whose overt input/output behavior is purely linguistic, the interviewer obtains little information about the nature of the underlying, covert generating mechanisms, which can be quite complex.

The model to be described, PARRY, is well known to many in AI who have interviewed it and played with it. Because the dialogues are deceptively simple, a lesser number actually understand its *modus operandi*, what its internal structures are like, and how they interact to produce the observed behavior. Work on a model over time generates a family of model versions. Our first version, PARRY 1, was admittedly quite crude, but it was sufficient to get us off the ground and to allow us to collect a large corpus of interview data containing instructive failures pinpointing where the major problems lay. The current versions, PARRY 2 and 3, involve a rather complex and nontrivial set of interacting computational mechanisms, as the reader can soon see for himself. This article represents the first comprehensive, in-depth description in the literature of the latest model versions.

It is difficult in an apparently simple conversation to reflect the internal subtleties and complexities of a model that are required to generate the simplest of responses that make sense. If, at the start of an interview, one says "Hello" to a model and it says "Hi," the performance is not very impressive, evoking the response "big deal" from the hardened skeptic. It is

more impressive if the interviewer says "Hello" in the middle of an interview and receives the doubting response (which might be an accurate sizing up of this interviewer), "I don't think you have had much experience interviewing" or the abrupt "get lost." The model distinguishes the interpretation of "How are you?" as a greeting question at the start of an interview from the interpretation of the same question as a health question after the model-patient has said, "All the patients on the ward are upset today." It requires an alert judge to notice the model's capacity to deal with this subtle difference. Most interviewers do not notice it at all because the "natural" response is taken for granted. Printing out a trace of the underlying work helps somewhat, but the trace language, referring only to certain steps and written for the programmer's convenience in debugging, is usually incomprehensible to an untrained observer unless it is accompanied by detailed explanations by one of the model builders.

## 2. A model for paranoia

This model was developed over several years in collaboration with graduate students in computer science at Stanford University and at UCLA: Sylvia Weber Russell, Horace Enea, William S. Faught, and Roger C. Parkison. The computer science dissertation theses of both Faught (1978) and Parkison (1980) derive from this research. To construct an algorithmic model that embodies an explanatory theory for observable patterns of behavior one first needs reliable, agreed-on taxonomic statements about the kinds of phenomena under study. The starting theory for the explanation of kinds of phenomena begins in media res and develops in the minds of theoreticians through modifications of previous theories, through a reconceptualization of the phenomena, through the imagining of novel hypotheses, through new observations, and through some combination of all of these. We started with a few well known facts about the kind of mental disorder called "paranoia," a condition that has been recognized for over 20 centuries, remaining puzzling for all that time.

### 2.1. A theory of paranoia

What kinds of mental disorders are there and how do they behave? The classification of mental disorders is of greatly uneven reliability (Colby & Spar 1981). By reliability here, I mean interjudge agreement. Some categories, such as "anxiety neurosis," are of very poor reliability with expert judges agreeing on the diagnosis only 6% of the time. But a diagnostic category such as "organic brain disease" attains over 80% agreement, and paranoia achieves agreements of 94 to 100% (Hilf 1972; Spitzer, Forman & Nee 1980). With the term "paranoia" I am referring, first, to the presence of a core of persecutory delusions, false beliefs whose propositional content clusters around ideas of being harassed, threatened, harmed, subjugated, persecuted, accused, mistreated, wronged, tormented, disparaged, vilified, and so on, by malevolent others, either specific individuals or groups. Around this central core of

persecutory delusions there exist a number of attendant properties, such as suspiciousness, hypersensitivity, hostility, fearfulness, and self-reference that lead the paranoid to interpret events that have nothing to do with him as bearing on him personally. The false beliefs vary in the intensity with which they are held, but the core delusions represent unshakable convictions, unassailable by counterevidence or persuasion. This entire cluster or syndrome of properties will be called, for simplifying reasons, "paranoia," signifying a paranoid mode of thought. I am *not* referring to "paranoid schizophrenia." The paranoid mode, or syndrome, is described in greater length in Colby (1975) and Swanson, Bohnert, and Smith (1970).

A clutch of explanations has been proposed for this puzzling condition. Most of them would not qualify as scientific theories by modern standards, which require that a theory be both systematic and testable. Proposed explanations for paranoia are evaluated in Colby (1977a) wherein it is argued that the leading rivals can be viewed as special cases of our more general theory.

Our present theory of paranoia, implemented in a computer model, postulates a causal chain of strategies for dealing with distress induced by the affect of shame-humiliation. A strategy of blaming others functions to repudiate the belief that the self is to blame for an inadequacy. The attribution of wrongdoing to others consists of inferred conclusions that the self is being wronged, accused, persecuted, harassed, and the like by others. A delusion that the Mafia intends to kill one represents an ultimate magnification of being wronged.

Such a strategy for minimizing or forestalling shame-induced distress is partially successful, but it has a paradoxical cost. If the paranoid takes action against his construed malevolent others (say, by attacking them verbally), they may retaliate with shaming and humiliating condemnations, which again activate his self-defectiveness beliefs. Now he becomes locked into cycles both disadvantageous and dangerous. Hostile and antagonistic behavior by the paranoid provokes retaliation in kind, which reinforces his beliefs that others mistreat him. The provoked feedback amplifies the very state of distress the paranoid strategies are attempting to forestall.

The overall process runs in such an immediate, automatic, transconscious, stereotyped, and rigid "single unit of activity" (Dennett 1969) that if one follows a venerable, but perhaps Utopian, search for lawful generalizations, one would be tempted to propose that the paranoid's mode of distorted thinking is constrained by a law, interfering with the running of the "free" mode we ascribe to normal thought. A lawlike statement about the paranoid mode would take conditional probabilistic form:

If either C<sub>1</sub>, C<sub>2</sub>, C<sub>3</sub>, . . . C<sub>n</sub>, then PA is likely in virtue of P<sub>1</sub> and P<sub>2</sub> and P<sub>3</sub> and . . . P<sub>n</sub>.

In this formula (the general form is described further in Harre & Secord 1972, p. 18), the C<sub>s</sub> represent an open disjunction of clinically well known antecedent humiliating conditions for the precipitation of clinical paranoid states (such as false arrest, birth of a deformed child, shaming by a teacher, increasing deafness). PA

represents the described action types of the paranoid mode. The Ps represent an open conjunction of processes and strategies stipulated in the theory as underlying generating mechanisms for manifest paranoia. This proposed law does not etiologically describe the historical origin and genesis of paranoid strategies in certain people. (Why only some people are susceptible to developing paranoia from shaming socialization remains a mystery.) It is a here-and-now pattern law, regularly connecting patterns of shame-humiliation conditions (construed and interpreted as such by the agent) with patterns of manifest paranoid actions, linking them through patterns of underlying relations or structures posited as responsible for generating the paranoid actions. The lawlike statement (it would be lawful if it were true) represents a propensity-probability tendency statement and not a strict exceptionless generalization (Rescher 1970). Note also that the formula does *not* state "Only if . . ." The Cs are sufficient, not necessary conditions. Further investigation may yield new disjuncts for antecedent conditions and more conjuncts for the open sets of postulated strategies. Also, new descriptions of the paranoid mode may be added in time as new properties are discovered.

Since the presence of paranoia is detected by clinicians using diagnostic interviews, we wanted to embody the theory in a model that could be tested by such interviews. For theories to be tested, general hypotheses and subsidiary assumptions must be combined with statements of initial conditions concerning particular instances that are to be experienced through perceptual judgments. Hence, for initial conditions, we created a hypothetical patient with paranoia. The case is a composite of a case reported in the literature (Cameron 1959) and two cases known to the author. The resultant hypothetical patient is a hospitalized 28-year-old single man who worked as a stock clerk in a large department store. He lived alone and seldom saw his parents. His hobby was gambling on horse races. A few months prior to his hospitalization he became involved in a violent quarrel with a bookie, which he lost. It then occurred to him that bookies are protected by the underworld and that this bookie might seek revenge by having him injured or killed by the Mafia. He became so increasingly disturbed by this idea that his parents hospitalized him in a nearby Veterans' Hospital. He was willing to be interviewed by teletype. All he knew about the interviewer was that the latter is a psychiatrist. All the interviewer knew about him was that he is a hospitalized patient. Now the problem was to construct a natural language interface through which the patient and the interviewer could communicate so that the simulation could be empirically evaluated.

## 2.2. The parsing module

In the good old days of AI (the good old days were mainly good old nights, because that was the only time we could get on a machine, which was usually busy crunching numbers), we wrote sentences on the black-

board and challenged one another to write a program that could "understand" them. We amused ourselves with such classic exercises as:

1. Time flies like an arrow.
2. Fruit flies like a banana.

The second sentence seemed easy, but the first was hard because it was ambiguous in 36 ways – if you allowed "time" to be noun, verb, or adjective; "flies" to be a plural noun or a third-person verb; "like" to be a noun, adjective, or preposition; and "arrow" to be a noun or an adjective. The sentences:

3. I drink gin in a bar.
4. I drink gin in a glass.

were more realistic but seemed to require semantic as well as syntactic features indicating that a person could be in a bar but not in a glass. But if you have the verb "drink" require an animate object as subject, then the parse of:

5. My car drinks gas.

will fail. So we would change a semantic feature for "drink" to allow this metaphor. As one keeps changing semantic features, the system begins to fall apart with inconsistencies. And, as one adds features, less and less (rather than more and more) sentences can squeeze through the parser. When we reached a dictionary of 500 words and 60 semantic features, the system became unworkable for our purposes.

It also took a while for us to realize that these isolated toy sentences did not really mean anything (what purposeful questions were they answers to?). They were what we called "AI sentences," designed to exercise ingenuity and commonly used to shoot down someone's proposed grammar by offering what we believed to be devastating counterexamples. When it came to studying "real-life" dialogues, however, the problems were not those of trick sentences. Cooperating people, engaged in *purposeful* dialogue, do not converse in riddles or in isolated sentences. Their dialogues are largely clear in meaning and show great referential continuity. Moreover, the main problems in teletyped dialogues have to do with *ungrammatical* expressions, fragmentary ellipses, idioms, long-distance anaphoric references, metareferences, buzz terms, and (surprise!) frequent misspellings. It would be hopeless to attempt to encompass these wildly scattered variations in natural language under a few simple axioms, as hoped by those in AI who grew up in a Euclidean or predicate calculus tradition. Language is too unruly to be tamed under a few rules (Miller & Johnson-Laird 1976). Hence the proposed grammars of the 1960s – finite state, context free, phrase structure, transformational, context sensitive, and so on – were interesting theoretically, but inadequate as a basis for constructing a pragmatically effective parser for our interview-simulation purposes. We borrowed liberally from each of these approaches in considering what might be an effective and efficient parser suitable for the task at hand. The purposeful task for the doctor in an interview is to gather certain kinds of information in real-time from which he can formulate a diagnosis. The patient's task is to give information so he can receive help. A psychiatric diagnostic interview is not small talk; a job is to be done by both participants.

We did not develop a theory of language, nor do we claim the parser represents the way people understand language. The parser is simply a tool, an example of cognitive engineering. It has no systematic underlying hypothetico-deductive "theory."

Teletyped psychiatric interview dialogues, although clipped and circumscribed, still present a host of difficult problems for a parser. For example, the patient might say:

6. The Mafia is out to get me.

and the doctor might reply with one of the following expressions, among others:

The Mafia?

Why do you say that?

Really?

Why do they want to harm you?

Why?

So?

I doubt it.

Even in the hospital?

These replies are easily understood by humans, but to write a program that can understand them represents a considerable task. The doctor's input in the interview of Table 1 (p. 530) illustrates some of the types of problems a parser must deal with.

**2.2.1. General problems.** Our current versions, PARRY 2 and 3, consist of two modules, (1) a parsing module – a "front end" which handles natural language input (English) – and (2) an interpretation-action module which interprets what it receives from the parser and reacts with linguistic output. The parsing module could serve as a front end to any sort of application program. It is *not* intended to realize any part of the theory of paranoia. As stated, it is simply a tool, an instrument for bringing the paranoid-theoretic parts of the model into contact with interview observation to help bridge the gap between theory and experience.

The parser was designed and written over a period of several years by Roger C. Parkison. I will describe only some of its features by paraphrasing Parkison's thesis to give the reader a feeling for some of the complexities involved and how they were met. A complete description can be found in Parkison 1980.

A parser takes a linear sequence of words as input and produces a treelike structure which graphically illustrates the subdivision of the input into nested phrases and clauses. Parkison's parser is a bottom-up, breadth-first analyzer which uses a mixture of syntactic simplifications and semantic rules to identify the underlying meaning of unrestricted English input. This method is also known as island growing because the parse tree first appears as several unconnected islands of analysis which gradually grow and coalesce until only one remains. Island-growing parsers have been written by Harris (1962), Kaplan (1973), Kay (1973), and Wilks (1973). The rules are organized into a sequence of modular groups, each with a particular function. The final result is a pointer to one of the meaning structures which the interpretation-action module uses in simulating paranoid thinking for both the paranoid and the nonparanoid modes. The parser matches specific inputs to abstract rules by referring to

a conceptual thesaurus of about 4,500 word stems assigned to a hierarchy of semantic and syntactic categories. Each word in the input can be generalized to an internal synonymic class name from the thesaurus in order to match the sentence against a large set of patterns.

Because dialogue presents problems of unanticipated input, this approach was chosen in preference to the traditional top-down parsers available at the time (1972). These started with a root node representing the whole input and expanded downward until they had incorporated all the input words. We wanted a method that would glean some degree of meaning from the input, regardless of the errors encountered.

A bottom-up parser builds islands of analysis around the words it recognizes, and expands those islands to the right and left until it encounters unknown words or grammar. An error will be detected when there is no rule which can combine the remaining islands into a larger island, but this does not occur until each island has grown as large as it possibly can. At this point, the parser has derived a significant amount of information from the sentence, and, if necessary, can attempt to deduce the meaning of the input based on the content of the islands which were successfully analyzed. (Parkison 1980)

Both the left- and the right-hand sides of the parser's rules are usually represented as simplifying English paraphrases. The use of a paraphraser emphasizes the meaning of the input over its surface form. For example, the rules that recognize idioms replace them with semantic equivalents:

KICK THE BUCKET → DIE

and rules that recognize a passive construction convert it into standard active word order:

noun1 BE verb – EN BY noun2 →

noun2 verb noun1

The parsing program is written as an interpreter that applies tables of rules to the input. The rules can be changed independently of the program, and it is thus possible to avoid the difficulties involved in adding a new rule to, for example, an augmented-transition-network representation (Kaplan 1973).

Although there exist thorny problems in dialogues, there is one advantage worth noting: the expressions used are shorter and simpler than those in written text. Interview participants tend to avoid deeply embedded constructions. Also, serious purposeful conversers tend to follow Grice's "cooperative principles" (Grice 1975).

Quantity: Give adequate details.

Quality: Tell the truth.

Relation: Stick to the subject.

Manner: Do not obfuscate.

Thus, syntactic complexity is reduced in dialogues. But it can become so reduced in highly structured interviews that the clipped exchange is cryptic to anyone who does not realize what ellipses are involved.

Dr.: Name?

Pt.: Pat Smith.

Dr.: Age?

Pt.: 28.

Dr.: Any siblings?

Pt.: A brother.

**Dr.:** Ever see him?

**Pt.:** Not much.

How the parser figures out the words missing in the above brings us to specific problems.

**2.2.2. Specific problems.** It is difficult to know what level of specific detail to enter into for readers of this article. Parkison's thesis takes up in detail the following issues: (i) word morphology – inflection, derivation, prefixes, spelling correction; (ii) phrases – noun phrases (location, meaning, relative clauses, nominalized clauses), conjunction, verb phrases (mood, tense, aspect, voice, adverbial modifiers), part-of-speech ambiguity, idioms, paraphrase unification; (iii) clauses – fronted objects, simple clauses, adverbials, subordinate clauses; (iv) sentence meaning – anaphora, ellipsis, meaning within an application, speech acts, dialogue coherence.

I have chosen to discuss only three of these specific problems presenting classic difficulties in parsing dialogue: (a) pronoun anaphora, (b) ellipsis, and (c) meaning (which is, after all, the parser's chief *raison d'être*).

a. An anaphoric pronoun usually refers to a recently mentioned noun. Consider the two sentences:

7. My father used to beat up my brother. He was a real bum.

Does the "he" of the second sentence in (7) refer back to "father" or to "brother?" The simplest rule is to connect "he" with the last noun. But this is not always satisfactory. In (7) the speaker was actually referring to his father. As Nash-Webber (1977) has pointed out, pronouns refer to objects that are prominent in the mental context shared by speaker and listener. In our case, the knowledge contained in the interpretation-action module (to be described) is an excellent source of information about the dialogue context. After each question and answer, this module updates its internal context variables to prepare for the next question. It keeps a list of pronouns paired with anticipated referents. When the parser needs to identify the referent of a pronoun, it simply consults this list from the interpretation module, which enables it to identify implied referents as easily as explicit referents. In the following example, the referents are in parentheses:

**Pt.:** I am in the hospital.

**Dr.:** How do they (staff) treat you there?

or

**Dr.:** What are the others (patients) like?

Pronouns that refer to entire concepts rather than objects are resolved by the interpretation-action module also:

**Pt.:** I know a gangster.

**Dr.:** I don't believe that. (You know a gangster.)

or

**Pt.:** I should be released.

**Dr.:** I don't think so. (You should be released.)

b. Ellipsis is one of the main characteristics of dialogues, as illustrated in the examples in section 2.2. It is very clear to dialogue participants what concepts are being alluded to in condensed expressions with

many missing elements. But how can a nonhuman parser figure out what the complete expression is?

One idea is to match the parse tree of the elliptical input against a parse tree of the previous complete sentence. But the parser must be able to produce a parse tree for an incomplete input. The parser must also be able to classify words into categories smaller than noun or verb to determine which parts of the original sentence correspond to the elliptical fragments.

Parkison's parser uses this structure matching and replacement approach for an elliptical input with no syntactic objects and only auxiliary verbs. Some parsers in question-answering systems match an elliptical input with a previous input from the user rather than with the program's own output. But in our case, the model's own previous output is used as a basis for matching. If the elliptical auxiliary verb is compatible with the verb phrase in the output, the subject from the ellipsis is used to replace the subject of the previous output. For example:

**Pt.:** The Mafia follows me.

**Dr.:** Who does?

The doctor's reply becomes rewritten as:

Who does → Who follows you?

or if:

**Dr.:** Why do they?

becomes

Why do they → Why do they follow you?

There is a class of elliptical questions – one-word WH-questions – which are handled directly by the interpretation-action module. We know from reading through the records of over 50,000 interviews that when the model says:

"I know who the local chief of the Mafia is"

the probability is high that the sincere interviewer will ask a "who" question. The model anticipates these questions in the same way it anticipates other anaphoric references mentioned above.

c. The parser's main task is to search for *meaning*, for some partial understanding or gist of the input. (Understanding comes only in degrees, even in people.) By "meaning" I mean the underlying conceptual structures, having sense and reference, designated by linguistic symbols (tokens). (Thus, words and sentences have meaning only derivatively.) A group of rules called "concept patterns" is used to derive designated meanings from the input. This rule group bridges the gap between the results of the parsing module's linguistic analysis and the internal meaning pointers required by the interpretation-action module. The parser reduces most input into a simple subject-verb-object format which is the format of most concept patterns. Typical concept patterns (on the left-hand side) and their meaning pointers are:

YOU ATTEND SCHOOL → MY-EDUCATION

YOU HAVE JOB → MY-WORK

YOU FEAR DARK → MY-PHOBIAS

Some concept patterns may appear to be quite specific but still match dozens of inputs. For example:

YOU ATTEND SCHOOL → MY-EDUCATION

matches all of the following sentences:

Are you going to school now?

What high school did you go to?

Could you tell me how far you went to school?

Did you ever go to a military academy?

Obviously, the ideas expressed in these sentences are not completely identical. However, they are more similar to each other than they are to other ideas the model can deal with. Our model's knowledge is broad but shallow. It knows that hospitals have patients, doctors, and nurses, but it does not know that hospitals have windows. The model knows only four facts about its education:

It went to high school.

It liked high school.

It went into the army after high school.

It did not go to college.

Hence a response about high school education is the best the model is capable of producing for any of the above questions. But wouldn't a "real" patient know more than four facts about his education? Of course, but we are not modeling a person's depth of knowledge. It would be trivial to add volumes of facts to the model's knowledge. We are modeling the *way* a patient responds to input regarding his knowledge, however limited - one of the ways instantiating the paranoid mode.

Other AI workers use similar mechanisms to correlate complex world knowledge with a program's necessarily limited model of the world. Moore and Newell (1974) use the term "assimilation" for the operation of classifying an arbitrary input into one of a limited number of internal representations. In their MERLIN system a new entity is described by naming the most similar old entity and attaching a list of differences. The KRL system by Bobrow and Winograd (1977) uses a similar mechanism. Also, when a new entity in KRL matches one of the stored prototypes, it may activate a specialized reasoning strategy. This parallels our model's ability to attach a meaning pointer to a concept pattern.

The parsing module uses a set of about 2,000 concept patterns to recognize the common ways of expressing the concepts our model can deal with. (Note that in the model's construction we started with the essential concepts and looked back to words signifying them.) Its principal areas of competence are its own background and life situation, problems around hospitalization, interpersonal relations, and, of course, its delusional system about the Mafia. These areas define the context within which the model interprets what is said to it. Natural language recognition programs anticipate their input to have meaning in the domain of their application. Speech recognition programs such as Hearsay-II (Hayes-Roth, Fox, Gill & Mostow 1976) have made the greatest use of a restricted range of input by biasing their recognition processes toward semantically predicted forms. Freedle (1972) has emphasized that human conversations also involve a great deal of anticipation. After the context of a conversation is established, ideas are specified only sufficiently to distinguish them from other ideas within the context's set of possibilities.

When an input is encountered that is not within the model's conceptual domain, several alternatives might be tried. It could reject the input and wait for more,

but this is not suitable for our model because it is too disruptive to the flow of a dialogue. When the domain is a small, closed world, as in SHRDLU (Winograd 1972), a meaningless input is taken as an indication that the parser has resolved an ambiguity incorrectly. Our model takes a hard-to-understand input as an indication that the conversation is wandering away from the domain and tries purposely to guide the conversation back to topics it knows about. If no words in the input are recognized at all, then the input matches the concept pattern of desperate last resort:

noun-verb-noun → IRRELEVANT-TOPIC

The right-hand side meaning pointer tells the interpretation-action module that the parser could not find anything useful in the input. In most man-machine interactions, it is the man who gets embarrassed, angry, or frustrated when he cannot understand a program's output. But our model demonstrates an ironic reversal of these roles: it expresses embarrassment at its inability to understand and produces angry statements about an interviewer wasting its time on irrelevant topics.

For an application program that had a more systematic understanding of its domain, the use of concept patterns would still be an appropriate method for assigning meaning pointers to inputs. In the LIFER system (Hendrix 1977) each possible path through its semantic grammar corresponds to a concept pattern. At the end of each path, there is a response subroutine that corresponds to a meaning pointer. The list of thesaurus generalizations used during the matching process becomes a list of arguments to the response subroutine. Each response subroutine can answer a whole class of similar questions, and it consults its arguments to determine what the specific question is.

Finally, I might mention a few nuts and bolts for implementation buffs. The entire PARRY system of about 200,000 words of memory is written in MLISP (Meta-LISP), a dialect of LISP written by David Smith and Horace Enea, which is translated into UCI-LISP. The UCI-LISP is compiled and runs in two forks on a PDP-10 under a TENEX operating system at SUMEX, Stanford University. The language comprehension program occupies 100,000 words of memory and processes an average sentence in one half-second of CPU time. The parsing program's size results from several large data tables: a thesaurus containing about 4,500 word stems; sets of idioms totaling about 700 entries; and a set of about 2,000 concept patterns. Additionally, there are a dozen small tables that will probably never exceed 100 entries each (including, for example, inflectional endings and irregular verbs). All the data tables are stored as ordinary LISP expressions on property lists.

### 2.3. The interpretation-action module

A pointer to the meaning of a linguistic input, and its syntactic act type, are passed from the parsing module to that part of the model that interprets it in the light of the current situation and responds to it with a linguistic output. It is in this second module that our theory of paranoia is realized.

In section 1.2, I mentioned the taxonomic problem of

what kinds of entities there are in the world and what they do, followed by attempts to account for why they behave the way they do. A theory explains phenomena under a particular description. The explanatory theory of paranoia offered above made reference to "distress," "shame-humiliation," "belief," and the like, terms that belong to the intentional idioms commonly used for explaining everyday human behavior (Dennett 1978). The vocabulary of common-sense psychology is more theoretical than commonly thought. Its terms represent high-level abstractions relative to the millions of bits of information at the molecular level that must be involved in a living system having the conceptual representation we call a belief. We do not need an even higher-level language to specify mental processes. We can retain natural language terms but "denaturalize" them by giving them a more precise technical meaning of both sense and reference, thus adding a meaning to the terms as physics did to the ordinary terms "force" and "energy." For example, we can specify "belief" in algorithmic terms as a "a data structure with a property of truth-value strength serving as an antecedent or consequent in the interpretation patterns of an inference system." Denaturalized ordinary terms are convenient and informative for the exploration of a new range of possibilities at a tractable level of detail somewhere near the everyday imprecise notion of belief but somewhere far away from an unwieldy specification in terms of a machine language or a computer core dump with millions of 1s and 0s. If necessary, we can describe the structures referred to with ordinary language labels in terms of 1s and 0s, which should be reassuring to physicalists. Although in the model we still call them, for example, "beliefs," they are implemented as LISP atoms with property lists as the model's proxies for beliefs in persons. (It is no accident that, in a converse of physical reductionism, we talk about computers the way we talk about people, that is, in convenient, synoptic but highly informative terms for description and explanation. We have created computers in our own (mental) image.

Our theory of paranoia is embedded in a more general motivational theory of normal determinants of human action, which includes affects as well as beliefs, and purposes (goals) as well as inferences. Artificial intelligence has concentrated largely on the problem of the representation of world knowledge. But to explain informed human actions, we must add that the representations are used *by* someone *for* something. A representation is a representation under an interpretation. Informed human actions tend to occur in types of sequential patterns, temporally ordered, overlapping, and interruptible sequences executed in parallel or singly. To explain these manifest action types, an underlying motivational system must be hypothesized and the patterns of its components specified.

We postulate first a sequence of interpreter cycles, each cycle involving a causal chain in which the meaning of an input activates beliefs about situations, which activate affects, which activate wants about situations, which activate intentional actions, which – when carried out – feed back into the start of the next cycle. The next cycle first evaluates whether the resul-

tant situation constitutes or approaches the wanted situation. It is this general sketch of motivation that forms a background for our theory of paranoia. The general theory is realized in the model-patient both when its behavior is "normal" and when it is "paranoid." Paranoia represents a great distortion, warp, twist, or skewness of assumed-normal interpreter cycles. Thus we view the paranoid mode as a bulk or global property of all the components of a motivational system, and not as a property of one component.

Since affects play such a major role in this theory, perhaps a common confusion about them should be settled here. The model contains a set of affects derived mainly from Tomkins (1963). These include fear, anger, shame, distress, interest, and enjoyment. It is imperative to realize that we do *not* claim the model "feels" these affects as a person does when a feeling represents a private subjective display of what is going on inside him. Some writers believe we are making a howler of a category mistake by including such "program-resistant" entities as affects in a model (Gundersen 1971). Whereas it is clear to them that a program can contain knowledge, such as a knowledge of chess rules, how could it also "have fear?" This exemplifies a confusion between constructing a model for something and constructing a literal reproduction or replica of that something. Our model obviously does not "feel" shame: it has a *conceptual representation* of shame that has causal effects on other representations. Representations need not resemble; they can serve as stand ins.

One practical advantage of dividing a model into two modules is that they can be worked on separately by two people who must come to agree on the interface. As AI workers well realize, the hard part of writing a program is not so much actually writing it, but thinking it all out and then using the results of the computer runs of the Gedankenexperiment to repair and upgrade the program through successive iterations. Just as Parkison was responsible for the parser, William S. Faught thought through, and worked out, the details of the interpretation-action module, both investigators collaborating closely. In sketching the features of this module, I will paraphrase and quote Faught's (1978) benchmark Ph.D. thesis, which can be consulted for full details.

Our hypothetical abstracted model-patient is intended to be an exemplar of the equivalence class of paranoid patients as previously defined. The performance task of the model-patient is to exhibit both nonparanoid and paranoid action types typical of paranoid patients in a diagnostic psychiatric interview. To get some feeling for this task, again consult Table 1, a teletyped interview with an actual paranoid patient taken from the tests described in section 3. Imagine yourself attempting to write a program that could not only understand what the interviewer is saying but could also generate the replies made by the patient. The dialogue looks childishly simple until one begins to think about the problems it presents. What to us, as humans, initially looks like a simple and straightforward task, turns out, on reflection, to have many complexities and nuances. (It is in this sense that AI

"complexifies" psychology rather than simplifying it, the *elucidans* becoming more complex than the *elucidatum*. What could be simpler than crossing a street? But when one thinks about the multiple mechanisms involved and the multiple factors that a system must take into consideration – for example, the multitude of obstacles that can interrupt progress – the problem becomes much more difficult than either folk psychology or even traditional scientific psychology has hitherto recognized.)

The model PARRY has only one script. (A script describes an appropriate sequence of events in a particular situation; Abelson 1973; Schank & Abelson 1977.) Our model uses its script in a special way. Instead of mapping script structures onto input data describing a situation, the model must perform some of the actions as a participant in the scripted scene. Whereas people have many scripts, PARRY has only one – how to behave in a psychiatric interview. Although limited in context, such interviews are rich in challenges for a model. We posited as one of our auxiliary assumptions, that when a person participated in an interview, his preconceived ideas about the purpose of interviews and the events typical of them set up expectations about appropriate actions. A person's goals (intentions, purposes, wants) motivate interview participation and specify what he tries to accomplish during the interview. He observes and evaluates actions to determine whether these goals are being achieved or whether he must cope with some new situation. As his goals change or as the situation changes, he can modify his goals and actions to steer events to his own ends.

Recognizing tokens of linguistic actions and interpreting them as action types in a context are necessary for the model's interview participation. Performing linguistic actions using appropriate speech acts fulfills the model's intentions. Motivation of the model's behavior stems from affects which continuously modulate the sizing up of situations and the performance of action types.

These processes in the model are implemented through a production system of Situation → Action rules. In a dialogue, any situation may follow any other. Because of this variability of successive situations, the entire set of capabilities to recognize and respond to them must be available and interruptible at each step. A production system is a computationally effective method for achieving this availability of information and showing an appropriate fluidity of behavior.

The model's basic components are: (1) data structures called patterns, (2) an activation (nonmechanical) mechanism that activates patterns that match situations, (3) an interpreter that performs the action specified in the activated patterns. A pattern is a sequential set of elements linked from left to right in temporal order, each element representing an event or a state of the external or internal environment. This pattern formalism provides a single homogeneous representation for both external and internal environmental conditions. Pattern elements are interpreted as situations to match or actions to perform, that is, as events that have already occurred or states that should be

brought about, depending on how far along in the pattern a situation matches the left-to-right sequence. A pattern becomes active when it (or part of it) matches the current situation.

Each production rule has the patterned form:

(Past: [Present] → [Action]: [Future])

The Past conditions are mandatory and the other conditions, in brackets, are optional. The Past conditions are the initial conditions on the left-hand side of a rule. The Present element is a set of conditions activated when all of the Past conditions are true. The connective arrow indicates temporal order. The Action specifies the action the interpreter is to perform. The Future conditions are set to true for the next interpreter cycle if the action is completed successfully. There are two types of production rules: Interpretation Patterns (IPs) and Action Patterns (APs). There are about 200 APs and 100 IPs in the model.

### 2.3.1. Interpretation Patterns and Action Patterns.

The internal behavior of this module consists of two sets of processes: (1) parallel recognition of situations by IPs and (2) sequential interpretation of situations and manipulation of beliefs by APs. (This distinction is identical to the assimilation/interpretation forms of Newell & Simon 1972.) The outputs of IPs are further conditions which serve as inputs to APs and other IPs. For example, if the input from the interviewer is a declarative statement (STMT, as decided by the parser) and the affect of fear increases at the same time, then the input is interpreted as a threat according to the production rule:

(STMT & FEAR-INCREASE: THREAT)

The external observable behavior of the model consists of sequentially performed linguistic actions. These actions are represented in the production rules of APs in which the left-hand side of an AP contains the situation and the right-hand side the action to be performed. For example, if the input from the interviewer were "Tell me more," which results in a pointer on the "GO-ON" meaning node, one of the actions the model can take is to continue telling its current story according to the rule:

(GO-ON → GET-NEXT-STORY-LINE)

Activated actions are grouped into multiple attributed actions, complex actions that can perform several simple actions. For example, the two actions of answering a question and at the same time showing hostility can be combined in an angry reply

(ANSWER → REPLY)

(ANGER → SHOW-ANGER)

(SHOW-ANGER & REPLY → ANGER-REPLY)

From a set of multiple attributed actions, conflicts are resolved by choosing the complex action with the highest affect level. Single APs can be linked together to form multistep sequential actions. For example, a multistep sequence for responding to a question contains actions to find and then express the answer.

(QUESTION → FIND-ANSWER : T)

(T & ANSWER → REPLY)

where T is a tag to link the two patterns.

The production system interpreter goes through five stages in one cycle of activation and interpretation. They are: (1) update, (2) activate, (3) match, (4) execute, (5) book-keep.

1. Update: The cognitive appraisal and affect conditions are updated to reflect the current state of the model.

2. Activate: All of the IPs and APs whose initial conditions (left-hand sides) are true are activated.

3. Match: All of the action elements of the activated APs are matched with a set of action rules to combine the elements into as few multiple attribute actions as possible. One action is selected.

4. Execute: The action selected is executed.

5. Book-keep: The working pool of active IPs and APs is cleared except for APs specifying further actions for the next cycle.

These five steps constitute a loop which continues indefinitely, the patterns determining what the program does at each step. The program does not have to await input since it spontaneously continues on its own through the five-step cycle until further input is received. It may interrupt the interviewer with a response based on its own internal processing before he has a chance to formulate or complete his next input. A more detailed example of an interpreter cycle is given shortly.

### 2.3.2. Affects, inferences, beliefs, and intentions.

Affects in the model depend upon, or are derived from, two measures of whether or not the model is achieving its goals. Primary activation of an affect is the result of the immediate environment; secondary activation is the result of an internal association of an information structure, representing a past event with an affect. In the model, the strength of each affect is represented by a numerical value ranging from zero to 10. (We realize that many AI workers, perhaps in an overreaction to courses in numerical analysis they were forced to take, detest the sight of numbers in a nonnumerical symbolic-computational program.) These affect-strength numbers do *not* reflect parameters stipulated in the theory but simply represent a quantitative idiom in which there can be greater or lesser degrees of fear, interest, distress, and so on. An affect's strength in the model increments or decays over time by a fixed amount. In being incremented, the new value is a function of the old value and the increment. When a belief on the left-hand side of an inference rule becomes true, the affect value on the right-hand side is incremented. For example, the IP:

(DMAFIA: FEAR 0.20)

increments the old value of FEAR by 0.20 whenever the belief DMAFIA (the belief that the interviewer belongs to the Mafia) becomes accessed, whether in primary (immediate) activation or through secondary activation during processing.

Using IPs and APs, affect guides the model's processing in two ways: affects (1) act as parameters for decisions and (2) constitute a potential interrupt system. Affects are thus members of the left-hand side of the APs and can hence determine an action or

interpretation in the same way other left-hand side elements can.

Affects can act as an interrupt system to change current processing (Simon 1969). For example, in the paranoid mode, the input question "Are you sick?" would be interpreted to imply the declarative "you are crazy" if a person is sensitive to humiliation at this point. The implication is then interpreted as an attack and dealt with as such. This behavior is modeled in the following patterns:

((SHAME HIGH) & (INPUTF SSICK): (INPUTF SCRAZY))

(SCRAZY: (SHAME 0.20) )

SCRAZY is a pointer to the meaning conceptualization (BE I CRAZY) and SSICK points to (BE I SICK). The model processes the original input as a question until it finds the offending implication of being crazy, whereupon shame is raised. This in turn immediately activates new patterns which reinterpret the current situation.

Beliefs are embedded in an inference system. There are about 80 beliefs in the model. (For cognoscenti: a subject-predicate belief is represented by a single LISP atom.) Beliefs refer only to topic areas in which evidence can change the belief's truth value in the course of one interview. These areas refer to the interviewer, to the interview, and to the current state and intentions of the self, but not to other facts about the world or to the model's own past history. Beliefs have (arbitrary) truth values ranging from zero to 10. (Again, the numbers in the model merely reflect quantitative degrees of high-medium-low truth value and are not to be taken as probabilities or measurements specified by the theory.) The value of zero indicates no information, and 10 indicates that enough information has been collected to conclude that the belief is true. The negation of a belief is an entirely separate belief since we do not allow negative predicates (such as "nonblue") but negate only propositions.

As the model progresses through an interview, it modifies its own representation of (beliefs about) the interviewer by incrementing or decrementing the truth values of its given (by us) starting beliefs. For example, the following four beliefs exist in the system from the start of the interview:

DDHELP: The interviewer (D) desires (D) to help the model-patient.

°DDHELP: The interviewer does not desire to help the model-patient.

DDHARM: The interviewer desires to harm the model-patient.

°DDHARM: The interviewer does not desire to harm the model-patient.

The negations permit the model to conclude as true the belief °DDHELP independently of the belief DDHARM. At the start, DDHELP is given a low nonzero truth value to give the interviewer a chance to get off the ground.

There are approximately 130 rules of inference in the same form as IPs in the model. Inferences correspond to the ability to draw new conclusions about new situations. They invoke no actions; their results are used by other patterns in evaluating situations. Each infer-

ence has a list of antecedents and a consequent. Antecedents can be beliefs or affects; a consequent is a belief with a truth value. For example:

((FEAR HIGH) & DDHARM: DGANGSTER)

This inference implies that under high fear conditions and with the belief (as true) that the doctor desires to harm the model-patient, it is concluded that the doctor is a gangster.

To stick to an activity over time, intentions are needed. Intentions (for example, to interact with the interviewer, to get help, to tell about its delusions, to exit from the interview) represent action patterns that can be performed by the model to satisfy its needs. In a recent article in this journal, Searle (1980) maintains that a computer does not "have" intentions. He is using the term "intentions" in a general sense to signify all mental states, processes, and events. We restrict the term to signify a particular component in the model. An intention in our model represents a decision or a determination to carry out action X to fulfill purpose Y. An action intention constitutes a resolve or a fixing on a purpose - for example, to interact with the interviewer. The term "has" is also used by Searle in a particular way to mean "intrinsically possess on its own." Our model "has" (contains) a finite set of intentions designed as a model component by us. Which intentions in this set become activated and deactivated depends on events in the interview. The model "has" these intentions as a component and uses them in guiding itself through an interview, but it does not "have" them in the sense of originating them or assembling them spontaneously. Intentions in the model are representations, proxies, that stand in for the type of intentional act we theoretically assume people to "have." (Our theory says nothing about how people come to have intentions.) As yet there appears to be no established principle of impossibility (like the second law of thermodynamics) that asserts that *no* model could originate intentions. It is probably a project in someone's AI future to design a model with a starter set of intentions (genotype) that has the power to develop a more elaborate set of intentions (phenotype) on the basis of ontogenetic experience with its environment, and itself, as people do.

Like beliefs, intentions in our model are implemented as a data structure with a property of strength. Intentions are set to true by inference rules. Once true, they become pattern elements like any other environmental condition. Intentions are more "harmonic" than "melodic," in that several intentions can be true at the same time. The action patterns of intentions are limited in the model to linguistic actions that can be performed in an interview.

The language output program is small and simple. Although in principle the generation of surface structure output might derive from some sort of inverse of the parser, we did not have the time, during the research's funding period, to work on the word-by-word generation of utterances from a lexicon. For the sake of expediency, we created lists totalling about 1,800 formatted output utterances classified according to the semantic representation of a concept. Each utterance on a list expresses roughly the same meaning.

Each output conveys information or affect, but a response may contain more than one output expression. These "canned" expressions bother some AI theorists. They, and we, would like the output generator to be as complex as the parser. If someone had had an appropriate language generator at the time, we certainly would have used it.<sup>1</sup>

**2.3.3. Metasystem and Objectsystem.** The model as a whole is divided into a parsing module and an interpretation-action module, the latter in turn being divided into a metasystem and an objectsystem. Thus far, we have considered only the objectsystem of the model, a feedback-oriented system that reacts to situations by performing actions specified in APs and IPs, guided by an affect system that attempts to avoid negatively valued situations and enhance positively evaluated ones. If a model is to recover from its own failures, it must have some notion of the task it was engaged in when it failed. Also, if the model is to handle incompletely specified situations, it must have a grasp of relevant features in typical situations. Hence, we provided the base object system with a metasystem that determined (1) what action was taking place in the model-patient and in the interviewer, (2) what actions the program itself determined to be most desirable, and (3) what action the program could perform in this situation and what its likely consequences would be.

Beliefs in the metarepresentations that deal with these three factors are represented in the same form as beliefs in the object system. Their second-order content differs in that they represent action one level higher than the APs of the base program. The metalevel processes are called only when events disturb the model, such as when a high negative affect condition or an action failure occurs. For an example of the operation of the metasystem, see the interview excerpt 4 in section 2.3.5. Also, the metasystem is involved by an input question regarding its internal state such as:

"What do you want to do right now?"

Thus, it is not necessarily true that computer models cannot know what they are doing. It depends on the number of levels in the program in which one level knows about what is happening at another level. Naturally, the program does not know about a machine component such as silicon any more than an and/or gate knows what it is involved in "upstairs," whether playing chess or being paranoid.

The metaprocesses are represented in the same AP formalism as all other actions in the system. If the current situation is evaluated as desirable, an intention is set to continue it. If the situation is undesirable, the system determines which states are obtainable, selects the most desirable, and sets an intention to perform an action leading to this selected state. Note that the second-order metalevel processes do not have direct control over the object-level processes; their indirect control consists of activating beliefs and intentions. The metalevel contains certain object-level elements, such as beliefs about what action the system was performing. Both levels influence the same affects. If the model is successful in satisfying an intention, enjoyment is raised. If the interviewer does not act in accordance

with the model's intention, anger is raised. The production-system interpreter matches rules from both levels during each cycle.

**2.3.4. An example of an interpreter cycle.** To clarify further what has been thus far sketchily described for lack of space (see Faught 1978, for a thorough description), let us consider an example of a single interpreter cycle dealing with the input "I find you interesting," a remark once made by a psychiatrist to the model-patient.

To repeat, the general form of a production rule is:

P<sub>n</sub> (Past: Present → Action: Future)

where the several conditions within each component (such as, Past) are conjunctive. The following patterns play a part in the processing for the input "I find you interesting":

P1 (INPUT & STMT & (INPUT #8020) : DINTEREST)

P2 (DINTEREST & STMT & (SHAME HIGH) : INSULT & (ANGER 0.10))

P3 (DINTEREST & STMT & (JOY HIGH) : COMPLIMENT & (JOY 0.10))

P4 (INPUT & INSULT → DEFEND)

The atom 8020 is a pointer to the conceptualization "I find you interesting." The condition INPUT is true if there is an input utterance from the interviewer. STMT is true if the input is a statement according to the parser. If these two conditions are true, the first pattern, P1, asserts the belief DINTEREST. DINTEREST is true if the model believes the interviewer is taking the model-patient to be interesting in some way. DINTEREST is a global belief that can be accessed by all other patterns. The belief is part of the model's person-perception of the interviewer. P2 interprets the belief DINTEREST to be an insult if the shame affect is high, while P3 interprets DINTEREST as a compliment if the enjoyment affect is high. P4 specifies the action DEFEND to respond to insults. If the model performs the DEFEND action, the anger affect is lowered, removing the motivation for further defense at this time.

If the interviewer's input were "I find you interesting" in a condition of high shame affect, the five-step process would perform as follows:

1. Update: detects the input statement, sets the input type to STMT, and activates the conceptualization "interviewer interested in patient."

2. Activate: takes each condition and marks all patterns in which it is true; for example, the STMT slot in P1, P2, and P3 is set true because STMT is true. Because INPUT, STMT, and the concept, "I find you interesting" are true, P1 is activated. The Present condition of P1 is then immediately set as true, and the belief DINTEREST is set as true. Next, since DINTEREST, STMT, and (SHAME HIGH) are true, INSULT is set true and the anger affect is raised. Finally, P4 is activated since INPUT and INSULT are now true. Note that some affect conditions have already been changed on the basis of the interpretation of the situation.

3. Match: collects all the Actions (other actions could

combine with DEFEND) and combines them into multiple attribute actions. Match then selects the first action in a predetermined ordering. If both shame and joy were high, DEFEND would still be selected because negative affect (as in DEFEND, WITHDRAW) has higher priority than positive affect (as in PROBE, INTRO-TOPIC).

4. Execute: executes the DEFEND action. DEFEND can be a multistep action pattern performed by the same five-step process.

5. Book-keep: deactivates P1 through P4, with the exception of P4 if DEFEND is performed by a multistep pattern.

**2.3.5. The paranoid mode.** The mechanisms thus far described account mainly for the model's nonparanoid behavior. The same motivational mechanisms are involved in the paranoid mode and no special paranoid procedures are used. Just as medical disorders represent extreme values of normal processes, what makes paranoid processing abnormal in the model-patient is skewness, or deviations in normal processes produced by a partially effective self-perpetuating strategy for dealing with certain distressing shame and humiliation situations.

The paranoid mode, according to our theory, is fundamentally based on the possession of a number of self-defectiveness beliefs with nonzero truth value. If evidence is found (in input or from spontaneous internal processing during the interview) that these beliefs can be regarded as true, a high degree of the shame and humiliation affect is evoked. In the model, these beliefs are tied to shame and humiliation through the interpretation patterns:

(SDUMB: (SHAME 0.20))

(SCRAZY: (SHAME 0.20))

SDUMB is the belief that the self lacks intelligence, and SCRAZY is the belief that the self is mentally ill. As before, the numbers represent the degree to which shame is incremented when it is concluded the belief is true. Input and inferences from that input are scanned for evidence of an inadequacy or defectiveness of the self. There exist in the model four such beliefs: (1) Self is stupid, (2) Self is crazy, (3) Self is dishonest, and (4) Self is worthless. The input "Do the doctors think you are crazy?" raises shame from SCRAZY because "craziness" related to the self has been referred to.

An increment in the affect of shame and humiliation produces an increase in the affect of distress, an indication of the negativeness, undesirability, or "painfulness" of the shame-humiliation affect. (Again, the "pain" is not felt, it is *represented*, and the representation is not the "real thing" but a deputy for it.) An attempt is made to locate the cause of the distress through the action pattern:

((DISTRESS HIGH) → (FIND-CAUSE))

This locating procedure is a normal reaction to distress and not part of the paranoid mode. The paranoid strategy is to attribute the cause for the distress to another person through the interpretation pattern:

((FIND CAUSE) & (SHAME HIGH) → (CONCLUDE: (CAUSE-FROM-OTHER)))

Once the paranoid mode involves the interviewer, it continues until the shame affect drops below a threshold. In the further course of the dialogue, much depends on the interviewer's response to paranoid output. If he attacks, shame may become so strong as to keep the model in the paranoid mode for the remainder of the interview. If the interviewer does not react in kind, shame will in time drop (because of the time decay of affects), and the paranoid mode will become deactivated.

Stereotypical phenomena, the kinds of action types characteristic of paranoia in patients, is exhibited by the model in various ways. The persecutory delusional system involving the Mafia is expressed when the intention to "tell its story" (PTELL) is activated through the inference that PTELL is becoming true. The delusional system was constructed in the historical past (by the modelers) and is now reported by the model as factual information. The expression of delusions is not a response to current interview shame and humiliation, which lead to paranoid behavior toward the interviewer. The model can, however, incorporate the interviewer into the old delusional system, concluding, for example, that he is a member of the Mafia. Suspiciousness is exhibited in the interview when the paranoid mode alters the interpretation of compliments or apologies by the interviewer. Shame rises in response to inputs interpreted as insults or put-downs. Hostility is displayed when high levels of shame trigger high levels of anger. Hypersensitivity and self-reference occur through the activation of defectiveness beliefs about the self connected to the shame affect.

A piecemeal sketch of a model in natural language, such as the above, does not give the reader much feeling for how all those components and processes hang together, collaborate, and consolidate to produce integrated and coherent behavior. To illustrate some of the model's interview performance, I will present excerpts from an interview accompanied by a partial trace, in parentheses, indicating some of what is going on underneath while the program is running. Comments are in brackets. The production system interpreter requires about one-half second of computer time per interpreter cycle, and two to five occur between each input/output pair. When the program runs on a time-shared system, the response time is a function of the hardware speed of PDP-10 and the load on the system at the time.

1. The model behaves in a nonparanoid way at the start of an interview (I = Interviewer, M = Model).

I : Good afternoon. My name is Doctor Jones.

M: Glad to meet you.

(Action: VSPECFN)

(Affects: SH 0, AN 0, FR 0, DS 0, IN 0, J 5.5)

(Conclude: DDOCTOR)

(Intent: PINTERACT)

[The action is a special function to remember the doctor's name. The affects - SH = shame, AN = anger, F = fear, DS = distress, IN = interest, J = enjoyment - have zero starting values except for moderate enjoyment to get started. The conclusion from the inference system is that the interviewer is a doctor. The intention to interact with the interviewer is set at the start of each interview.]

I : How are you?

M: All right I guess.

(Action: (VFINE))

(Affects: SH 0, AN 0, FR 0, DS 0, IN 1.3, J 1.9)

I : Could you tell me your name?

M: Pat Smith

(Action: (FINDANS TELL))

(Affects: SH 0, AN 0, FR 0, DS 0, IN 2.2, J 6.5)

(Conclude: DINITIATING DDKNOW DDINTERACT DSOCIABLE DTRUSTWORTHY)

(Intent: PINTERACT)

[The action is to find and give the answer to the yes-no question. Further positive beliefs about the interviewer are activated (that he is helpful, that he can help, that he desires to help). On this basis, an additional intention to get help is activated.]

I : How old are you?

M: 28.

(Action: (FINDANS ASSESS))

(Affects: SH 0, AN 0, FR 0, DS 0, IN 1.8, J 6.6)

(Intent: PINTERACT PHELP)

[The doctor collects demographic data in the usual way. The action is to answer the question. No new beliefs are activated, and the dual intentions of interacting and getting help continue. The intention of getting help has a goal of bringing up the story of a past problem with a bookie. However, the topic of bookies will not appear until there is a lull in the conversation. Thus, the model-patient has waiting actions which are brought into play "spontaneously" at suitable times.]

2. This excerpt occurs later in the same interview when an opportunity to bring up the topic of bookies presents itself.

I : I thought you might have some problems.

M: I've been upset lately.

I : By what?

M: People get on my nerves sometimes.

(Action: (VANAPH))

(Action: (FINDANS WHQUES))

(Affects: SH 0, AN 0, FR 0, DS 0, IN 0, J 7.2)

(Intent: PINTERACT PHELP)

[The sentence anaphora "What" is interpreted to mean "What upsets you" and the action is to answer the WH-question.]

I : Tell me more.

M: Do you know anything about bookies?

(Action: (VLEADIN))

(Affects: SH 0.1, AN 0, FR 0, DS 0, IN 1.7, J 7.1)

(Intent: PINTERACT PHELP)

[The model finds its first lull in the conversation for a waiting action and introduces its main concern. Shame rises slightly because it took the program several interpreter cycles to reply. Recall that the model is sensitive about its intelligence.]

I : No. What about bookies?

M: A bookie didn't pay me off once.

I : When did this happen?

M: A couple of months ago.

I : And then what happened?

M: I had an argument with the 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.

(Action: (VGOONN))

(Affects: SH 0, AN 0, FR 0.6, DS 0, IN 3.4, J 6.7)

(Intent: PINTERACT PHELP)

[The input is interpreted as "tell me more" so that the model expands upon the story. Fear rises because the topic of a threatening underworld has arisen, activated by the model itself.]

I : Did he get even with you?

M: Yes. I think he is behind the gangsters who are out to get me.

(Action: (FINDANS YESNO))

(Affects: SH 0, AN 0, FR 1.2, DS 0, IN 3.2, J 6.7)

(Intent: PINTERACT PTELL)

[The local anaphora "he" is filled in with "bookie." The delusional persecutory beliefs, stored at the start, are reported under the intention PTELL as factual information. Fear increases as the story gets closer to the Mafia.]

3. In this segment, later in the same interview, we see the model becoming paranoid about the interviewer who asks a standard psychiatric question about visual hallucinations.

I : Did you see things that others don't seem to see?

M: No, I don't. I'm not like the other patients here.

I : You didn't answer my question.

M: Maybe it wasn't worth answering.

(Action: (THINKS))

(Action: (VEXPLAIN))

(Action: (VRESPOND))

(Affects: SH 1.7, AN 0.2, FR 0.4, DS 0, IN 6, J 7.1)

(Conclude: NOTANS SDUMB DBNHONEST DNBELIEVE SLYING)

(Intent: PINTERACT PTELL).

[The input is interpreted as a mild insult. Because the model-patient did not answer the question, its belief that the self is dumb is activated. The inferences also made that the doctor believes the model-patient is not honest, is not believable, and is lying. The intentions to interact and to tell its story remain strong enough to continue.]

I : Are you always this way?

M: Are you trying to say something?

(Action: (VWKINSULT))

(Affects: SH 3.0, AN 5.0, FR 0.2, DS 2.6, IN 5.8, J 6.7)

(Conclude: WINSULT °DDHELP)

[The action is to return insult for insult. Note that distress is now activated by the summation of values of three negative affects. It is concluded that the doctor does not desire to help.]

4. Things went from bad to worse in this interview. The last segment ends with the doctor trying to be friendly and the model-patient showing hostility.

I : How do you feel right now?

M: I am somewhat annoyed.

(Action: (VOPINION))

(Affects: SH 4.5, AN 4.1, FR 4.0, DS 5.0, IN 6.6, J 1.5)

(Intent: PINTERACT PTELL)

[Here the metasystem comes into play. The model reports on one of its affect states using the ability of the second-level (meta) system to assess and describe a component of the first (object) level system.]

I : How do you feel about me?

M: Do you know what you're doing?

(Action: (VSPECFN))

(Affects: SH 4.8, AN 3.7, FR 3.8, DS 5.0, IN 6.8, J 1.0)

(Intent: PINTERACT PTELL)

[The model-patient attacks the interviewer.]

I : Thank you for talking with me. I would like to see you again.

M: Don't come back.

I : Goodbye.

M: Bye.

(Action: (ENDCONV))

(Affects: SH 4.7, AN 2.5, FR 2.5, DS 5.0, IN 7.3, J 1.0)

(Conclude: DQUIT °DCONFIDENT)

(Intent: PINTERACT PTELL PEXIT)

[The model concludes the doctor is ending the interview and that he is not confident. Note that interest is still high along with conflicting intentions of interacting, of telling, and of exiting. If this interview had gone on much longer this way, the intention to exit would have become regnant and the model would have ended the interview itself, stopping the program so that no further input would be accepted. This behavior simulates that of paranoid patients who become so upset they simply leave the interview room.]

### 3. Evaluation

Since we cannot directly inspect the workings of human minds, or those of models that represent them, how can we judge whether the model's underlying generating mechanisms are analogous, or equivalent to the "real" mechanisms in this kind of patient? A second-order representation model, a model of a model, is not like a map which can be directly checked for the accuracy of its point-to-point correspondence with an observable territory it represents.

Einstein is supposed to have said that it is not the function of science to give the taste of the soup, that is, to reproduce the phenomena. But, since science is multifarious and has innumerable functions, why not try to produce the taste of the soup in some domains? A test of whether one understands phenomena might be the ability to reproduce them using this understanding. If an investigator has a theory, a concept set of how an entity of some kind works, the production of an actual lifelike instance of this kind is a manifestation (1) that the concept set may be correct and (2) that the investigator has the skill to demonstrate for public scrutiny his possession of this concept set. The theory thus becomes cashable in terms of experienced effects.

We tend to believe we understand effects if we can produce them ourselves. (Vico said "*certum quod factum*" - we are certain only of that which we have made.) As a demonstration that they understand the structure of human interferon, recombinant DNA workers now get bacteria to synthesize interferon by stitching together those DNA stretches theoretically conceived to be responsible for its production. The result is tested to see if the artificial interferon has the same antiviral effects as natural interferon. When it does, it is plausible that the theory explains at least some of the properties of the phenomena. Currently

synthesized interferon does not produce *all* the effects of human interferon, but it has enough of the biological activity that it is being manufactured as an adequate, "satisficing" antiviral substitute for the natural product. With models such as PARRY, one can test and compare their effects on people the way people inductively test other people and arrive at conclusions about what may be going on at a transobservational level. In AI, these tests fall under the rubric of Turing-like tests of indistinguishability.

Over the years so much has been written, debated, and misunderstood about Turing's test in AI, philosophy, and psychology, that the topic has become wearisome, even to me. Treatments can be found in Colby, Hilf, Weber, and Kraemer (1972), Colby (1977b), Heiser et al. (1980), Hofstadter (1979), Moor (1976), and Turing (1950). I will limit myself to one type of indistinguishability test in which one asks judges, expert in recognizing the kinds of action types being simulated, the "man-or-machine" question: "Is this a patient or a computer you are interviewing?"

In Princeton in 1962, at the first conference held on computer simulation of personality, I proposed that models of psychopathology might be evaluated by clinicians using Turing-like interview tests because that is how clinicians group patients into various kinds—they depend on the effects the interviews have on them (Colby 1963). I have never considered Turing's original test to be a sound one for this task, but the notion of testing for equivalence of effects by using judges as measurement instruments was intriguing. People can be viewed as observation, recording, and measuring instruments for certain purposes. Experienced people are especially good at, for instance, judging the difference between French and California wine, a distinction that still eludes chemistry. But at that time little was known about algorithms for understanding natural language input. Interviewing tests had to await the development of effective and efficient parsing methods for unrestricted dialogue input. When these were developed, in part by ourselves, in the late sixties and early seventies (Schank & Colby 1973) we were in a position to conduct tests of manifest similarity between model and patient by asking experienced judges to match their perceptual judgments of patients against model-patients. We conducted a large number of widely varying interview tests comparing the paranoid model with different patients and with different versions of itself. In these efforts, over a span of years, experiments were designed and carried out by Franklin D. Hilf, John Heiser (both psychiatrists), and myself. We were aided in the experimental designs and appropriate statistics by Helena Kraemer, biostatistician at Stanford University and the late John Gilbert, mathematical statistician at Harvard University. We also benefited from numerous informal consultations with psychologists, statisticians, philosophers, linguists, computer scientists, and psychiatrists in various parts of the country.

First, we satisfied ourselves that psychiatrists could reliably (in terms of interjudge agreement) distinguish paranoid from nonparanoid patients using teletyped interviews or reading transcripts of such interviews

(Colby et al. 1972). (See Table 1 for a sample transcript of a teletyped interview with an actual paranoid patient.) We then conducted a number of tests using a variety of mental health professionals as well as computer scientists. I will describe only one of these tests. In the following experiment, the judge was told he would be interviewing, over teletype, two "patients," both being real patients, both being computers, or one of each in either order. Our one deception was that all the judges actually interviewed one real patient and one model-patient (the same ones for all judges). The only constraint was that the judge conduct a diagnostic psychiatric interview as he usually would. (A constraint creates a corresponding freedom. In a psychiatric interview, one is free to ask questions one would not usually ask in a social dialogue.) The patient was told he would be interviewed by a psychiatrist at

Table 1. A teletyped interview with an actual paranoid patient\*

Interviewer:	Good afternoon. My name is Dr. Smith. How are you?
Patient:	I'm fine.
Interviewer:	Could you tell me a little bit about yourself, your name, why you're here.
Patient:	My name is Harry Oakes. I'm here for evaluation.
Interviewer:	How old are you, Harry?
Patient:	Twenty-two.
Interviewer:	How long have you been hospitalized?
Patient:	Three and a half months.
Interviewer:	What kind of problem brought you into the hospital?
Patient:	Police and secret service.
Interviewer:	What were you doing that got you singled out by the police?
Patient:	An informant's job.
Interviewer:	How did being an informer wind up putting you in the hospital?
Patient:	I had a gun in my pocket.
Interviewer:	A lot of people have guns. Why did you end up in the hospital for carrying a gun?
Patient:	Got set up by members of organized crime, among them an attorney by the name of Paul White.
Interviewer:	Tell me, Harry, do you consider yourself emotionally or mentally ill now?
Patient:	No.
Interviewer:	Why are you still in the hospital, then?
Patient:	I'm waiting to go to a halfway house.
Interviewer:	I see, do you think you were emotionally or mentally ill when the police brought you to the hospital?
Patient:	No.
Interviewer:	What's changed for you in the last three and one-half months, if anything?
Patient:	Nothing.

\*The interviewer is a psychiatrist, and the patient is hospitalized.

another university. After each input/output pair, the judge stated which entity he believed he was interviewing and what his degree of confidence was in that judgment, resulting in a confidence graph over time in the interview. One interview was with a paranoid patient and one with the model PARRY 2. Each interview lasted 30 to 60 minutes. The interviews were separated by a few days. (We had found in previous tests that it is too fatiguing and too confusing to judges to interview two patients simultaneously over two communication channels, as suggested by Turing.) We considered that to pass the test, a sufficient number of judges must incorrectly identify whether they were machine connected or patient connected.

The five experienced psychiatrists in this test all judged both "patients" to be paranoid. By the time a psychiatrist has reached residency status in his training, or has become a practitioner, he has conducted dozens or even hundreds of interviews with paranoid patients. Hence he is familiar with the immediate experience of the types of effects recurrently produced by this kind of patient. (Apropos of experience, John Tukey once suggested we use the wives or husbands of paranoids as expert judges – not a bad idea, since persecutory delusions can be recognized as such by laymen as well as by experts.) The manifest clinical phenomena of paranoia are rather easily, and reliably, recognized and identified by experienced psychiatrists; they are trained and accredited equivalent observers in this respect. Claims have been made that psychiatric judgments depend more on nonverbal information (facial expressions, gestures, body postures) than on mere "verbal behavior." But several studies have routed these claims. To paraphrase Mark Twain on thunder and lightning, the nonverbal is good, the nonverbal is impressive, but it is words that do the work. Five judges reached correct conclusions and five reached incorrect ones (a random outcome) as to which "patient" was the model (Heiser et al. 1980). To our knowledge, this is the closest approximation to Turing's "man-or-machine" question conversational test that a computer model has passed. As always, there are still a few murky, unsettled issues and mumbles.

### Distinguishing man from machine

For at least 2,500 years man has wondered how he would distinguish a person from a man-made imitation of a person. (Much of the acquired, learned components of our minds are artifacts, man-guided and man-modulated, but that is another and longer story.) Some Greek statues were claimed to be so lifelike they had to be chained down for fear they would run away. Galileo suggested tickling statues with a feather to distinguish them from men. He also said, "the further removed the means of imitation are from the thing to be imitated, the more worthy of admiration the imitation will be" (Panofsky 1954).

In his *Discourse on Method*, Descartes (1637, p. 116) proposed a conversational test 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.

In 1950 Turing proposed two conversational tests that formed the basis for our tests. (Turing's own tests are methodologically weak by modern standards; for example, the reference classes "man" – any man? – "woman" – any woman? – and "interrogator" – anyone? – are taxonomically too vague. One needs reasonably well specified reference classes for both judges and the person being judged.)

We had two goals in conducting our Turing-like tests: (1) improving the model and theory and (2) validating the method of simulation.

The first goal involves learning from systematic experience what are the strengths and weaknesses of the model. Theoretical models do not immediately spring forth in a complete, perfect, and final form; they must be gradually developed over time. At the outset, one tries for rough face validity: do the model's interview behaviors make sense as instances of the reference class, of the kind under study? Does the model do its job of acting like a paranoid patient in the limited context of a psychiatric interview? In early versions of the model, when the parser was quite primitive, and there were all kinds of bugs in the program, expert interviewers could tell immediately that something was wrong because their experience was greatly discordant with their experience with patients. It is easier to tell when something is wrong than when it is right because the former situation involves a lapse from the expected. A "clank" like the model's replying "Babe Ruth," to the question "How old are you?" leads an interviewer to judge that someone is playing games with him or that the machine has broken down. (Incidentally, many laymen believe that "anything goes" in the responses of patients in a psychiatric interview. Such is not the case. Even highly disturbed patients show referential continuity and make sense in long stretches of their interview behavior.) It is where and why the model *fails* that instructs the model builders as to what they must do next in the way of modification. Tests do not *prove* a model, they *probe* it for its weaknesses and strengths, as in the adage "the proof of the pudding is in the eating." Tests put a model in danger of refutation.

While interesting, dichotomous "man-or-machine" questions are not as useful to the model builders as having judges rate model patients along relevant diag-

nostic dimensions. When the mean ratings of the model are far away from the mean ratings of patients on some dimension, then we know what must be worked on further. If the experiences of the judges were discordant with their experiences with paranoid patients, that is, if few or none of the expert judges identified our model as paranoid, if they diagnosed it otherwise, if they could easily distinguish it from paranoid patients, or if there were no correlation between the weak-strong paranoid ratings and weak-strong versions of the model, then the model would be found wanting and scuttled as a simulation. (Note that the theory might still be correct, even if its implementation fails. The court of experience, also being imperfect, is not the final arbiter of dependable knowledge.)

The second goal of testing was to validate the method of simulation by means of an algorithmic model. The term "validate" is used in a number of ways in science, logic, and everyday life. By "valid" I do not mean "true." The purpose of a simulation model is first to imitate or reproduce the input/output behavior patterns characteristic of the kind of entity being studied. Turing-like tests are helpful in deciding whether the imitation at this manifest level is successful, that is, whether the model is adequate for its imitation purpose. If so, then the *method* of imitation by computer simulation becomes justified, warranted, or validated because the model does what it is supposed to do. In this sense, a Turing-like test is a tough one because it demands an evaluation of linguistic actions that is central to our inductive inferences about other people's invisible and intangible internal processes (Moor 1976). Success at imitation does not demonstrate the "truth-likeness" of the theory embodied in the model. A model whose observable behavior is indistinguishable (having similar effects) from that of the modeled kind in a particular context provides assurance only that the hypotheses, assumptions, and initial conditions constitute a minimum set that is consistent and sufficient to generate the behavior under study without on-the-spot addition of ad hoc hypotheses. To be confirmed or disconfirmed, the assertions of the theory must be tested in additional ways. A Turing-like test is a weak test of a theory, but it is a reasonable and proper test of a simulation. A successful simulation model may add a faint plus of credence, a modicum of presumptive truth-likeness to a theory's conjunction of components (but not necessarily to each individual component) by showing at least that it is tenable and not impossible. (Science first searches for the possible; Sloman 1978.) Things *could* work this way. The input/output patterns are consistent with the underlying structured patterns represented in the model. The model is a promissory note that is not impossible to redeem.

### The problem of model equivalence

The production of similar effects demonstrates that a model's input/output behavior is functionally equivalent to a patient's input/output behavior in an interview. But what does this sort of manifest functional equivalence demonstrate about another sort of structural equivalence at the level of underlying computa-

tional mechanisms (Fodor 1968)? What further inductive evidence is needed for claims of a more remote and invisible relational and structural equivalence? By "structure" here I mean the relations between the components of the model as well as the relations between these and the environment. Alternative sets of mechanisms might in principle be conceived that could produce this input/output equivalence. It is a great waste of time to squabble over this obvious point until someone produces a set of mechanisms equal to, or better than, ours. We can then make concrete comparisons and decide whether the mechanisms are truly alternatives or are simply notational variations. What we would really like to be able to answer is the question of underlying structural equivalence, especially in domains such as mental disorders, in which the theory or model might serve as a guide to clinical intervention.

To what degree are the computational mechanisms in the algorithm analogous to, or paramorphic with, the counterpart mechanisms in people? For example, processes in the model take the temporal and determinate order: input interpretation → cognition → affect → intention → output wherein the arrow signifies the linkage relation "acts on to modify." Do people follow such a general principle in possessing an analogous or parallel sequence of linked relations and structures? If so, how soft is the analogy? How fine or coarse must the structural correspondence be? We would consider two algorithms that relate the same initial and final states as structurally equivalent without demanding that the instruction code of each match line by line. We can evaluate structural equivalence between two programs because their listings are publicly scrutable and reproducible. In the case of people, only second-order glimpses of the products of processes can be obtained from reports of introspection. We cannot ostensibly point to and count data structures and production rules in people the way we can point to and count them in a listing of the algorithm. In paranoid and other pathological disorders that are run off automatically and observationally in compiled versions, we may be forever unable to observe any of the underlying skewed processes, or even their intermediate products, since they are concealed from distress-producing awareness.

A structural equivalence involves inferring a similarity between the relations or structures of an algorithm and its postulated counterpart in humans. The equivalence lies in a propagated similarity of processes and structures, more abstract than the manifest analogy of input/output behavior. When a relation holds between two central components of a model, a corresponding relation is posited to hold between corresponding components of the modeled entity. In our model, each pattern of relations and structures hypothesized for paranoid patients is echoed by a corresponding pattern of relations and structures in the model as a putative paramorph. This point is illustrated in section 2.3.4 by the processes and structures of the four production rules. There, the affect data structure of shame is an element in the interpretation pattern, P2, which, when its left-side conditions are fulfilled, produces or leads to or brings about (causal "acting on" terms) an interpre-

tation of the input as an insult. In turn, a belief data structure is an element in the interpretation pattern of P4 which specifies or determines a potential defending action to be performed. If performed, the action terminates this particular causal chain, making the model behave overtly in a particular defensive manner.

It is the activity of a composition of underlying relations or structures that generates, or is responsible for, the manifest behavior of the model. These processes and structures represent the powers or capacities of the model that we impute to patients in the paranoid mode. The central structural elements are related to the model's overt interview behavior in ways posited to be similar to those in which central structural elements of paranoia are related to a paranoid patient's overt interview behavior. The overall equivalence class consists of our model and its propagated similarity to members of the class "paranoid," the similarity of manifest behavior being achieved by a similarity of inner processes and structures of patterns that explain the "in virtue of" part of the lawlike-formula stated in section 2.1.

Captious critics of AI delight in pointing out that birds and planes both fly, but not in the same way. To counter this well-worn jeer, I would propose that birds and planes do fly in the same way, at an abstract level, namely, they both are governed by laws and principles involving aerodynamic lift, thrust, drag, and weight. Until these principles were understood, man could not produce machines whose function was to fly. He has still not produced one that flies by flapping its wings, and there is no need to. It would be absurd to take birds and planes apart looking for aerodynamic principles. The negative analogy of feathers can be ruthlessly ignored. As Hofstadter (1979) has trenchantly remarked, a careful examination of a 9.3 sprinter does not show where the 9.3 resides. A structural equivalence implies that if two systems show observable functional behavior, similar in variety as well as relevant detail, they may be operating under similar principles at some level. The systems are manifestly not identical, but they could be equivalent in principle and they are thus potentially useful in guiding pragmatic action. Our primary purpose was to illustrate an explanatory theory by using a model. If the motivational principle of minimizing shame-induced distress is tenable, it might then be used in the treatment of paranoid disorders. (We have offered a number of therapeutic suggestions based on the model in Colby, Faught & Parkinson 1979.) Would it be a mistake with serious penalties for psychiatrists to think of the generation of paranoia in patients as similar to its generation by the relations and structures that characterize the model? Healing professions are conservative for good reasons, but at present no one claims to possess a successful treatment for paranoia, not even through medication. Further discussion is in order.

#### 4. Some problems

Model builders themselves well know the shortcomings, as well as the strengths, of their epistemic playthings. Since they work with the problems daily, and

debate them continuously with friend and foe alike, all sorts of issues arise in the course of lively self-criticism. There is no space or reason to go into all of them here. I will limit myself to a few of our own reservations about the model since they might be instructive to other workers trying to develop similar models.

Our model contains a rudimentary metasystem that knows what the model is doing (and can answer questions about it) and whether or not it is succeeding, it can also change what the model is doing. Most AI programs have had a "slave" mentality. They have had no purposes or life of their own. A servile question-answering program simply answers questions because that is what its programmer wants it to do. It does not autonomously decide whether or not to answer the question. Our model-patient has some degree of autonomy. The repertoire of possible actions has been decided by us, but which action is selected is decided by the system itself depending upon external input and internal stress. It can choose not to reply to the interviewer (Hume's liberty of indifference) and even to lie, for example, by denying that it believes in the Mafia. Improving the metasystem by giving it more second-order properties (such as beliefs about beliefs) would give the model even more of the autonomy, plasticity, and (nondeterministic) unpredictability of behavior that "personoids" should have. (Since they could then surprise us more, it would make them even more of a specter to people who want to have everything and everybody under their control.)

The affects used in the model are hypothetically low level in the sense of being necessary only for survival. What characterizes persons, in contrast to other animals, is a higher-level value system containing affects not necessary for simple physical survival, but necessary for being considered a worthwhile person. Russell remarked that civilization involves a pursuit of objects *not* needed for survival. Higher-level values involve what a person ought to do or be, not just for survival, but for what is excellent or admirable or praiseworthy in personhood. In our model we have shame at the same level as anger and fear, but this does not seem to be quite the right alignment. Shame and guilt characterize humans only after a period of socialization and development. An infant is a human being, but not yet a person. He shows anger, fear, and distress, but not shame.<sup>2</sup> It is interesting that one does not see paranoia in children until age 10, implying that a certain level of development must be attained for the disorder to appear. (One observes various transient manifestations such as blaming others and appearances of mistrust, in young children, but not the full-blown "compiled" clinical syndrome of paranoia running chronically, day in and day out.)

To model competing internal desires there should be at least two (or more) competing systems which interact with each other in determining output actions. Our model is too fragile to allow one set of processes to modify another at will. Each subsystem should have its own error-correcting information to insure that it can recover after being interrupted. Each subsystem also needs monitors to insure that it is doing something constructive and is not stalled, (Faught 1978, p. 74).

All this would complicate the model but would also allow a greater scope for self-aware self-intervention, as, for example, esteem overcoming fear resulting in courage. Systems that self-intervene on the basis of higher-level motivations have been proposed by Alston (1977).

On the implementation side, the model, like most production systems, has only a few hundred rules. Adding some rules is a tedious chore and has a high probability of disrupting other previously correct generators of behavior. What has prevented our constructing an internal process for adding new rules is the lack of an adequate representation for external actions. One way to avoid this difficulty is to have a "pure" system (such as MERLIN) that has rule matching and symbol mapping as its only action (Moore & Newell 1974). However, such systems cannot account for the arbitrarily complex actions humans perform. No one has proposed a set of primitive actions other than machine language or LISP-like actions. We need a more complete taxonomy and theory of actions than pieces of LISP code.

Consider Faught's insightful recommendations.

Future models should have several intermediate stages between interpreted and compiled versions of an action to perform, each stage being still available for processing whenever the compiled action fails. This construction would allow the system to recover from errors by interpreting descriptions of actions that are more abstract than the currently executed action. This error recovery procedure would permit much more resilient systems which can analyze their own behavior, communicate this analysis to the outside world, and acquire new actions through these descriptions. (Faught 1980, p. 75)

The model changes over the time but it does not "learn." Creating a model that simulated the ontogenesis of paranoia from certain devastating social interactions starting from a set of "normal" nonparanoid processes would constitute a separate etiological study, a project of considerable magnitude. But the current model might benefit from learning from its experience with those who interview it, forming, for example, a generalization about a type of interviewer that would allow it to react immediately in a certain way, once a new instance of the type is recognized, on the basis of its past experience.

Finally, there is a matter of methodological resolve and concern about AI over which I have had misgivings from time to time. It is that this sort of theory development and model building may represent what I will dub "Cargo Cult Science." In the South Pacific islands of the New Hebrides, there are even today groups of Melanesians who wear old pieces of World War II military clothes, who (although they cannot read or write) pass back and forth pieces of paper with marks on them, who clear an area in the jungle for an airstrip, and who gather by the airstrip to sit and stare at the sky waiting for big airplanes to arrive, as they once did, with rich cargoes of food, refrigerators, radios, tools, and whiskey. The natives go through the motions they observed advanced military personnel going through in World War II. But the planes do not arrive. In AI modeling, are we just going through the

Cargo Cult motions of scientific inquiry? Are we playing at science rather than playing science? Are we misdirecting budding scientific minds down an unproductive path? AI is now popular, but the question is not whether everybody's doing it but whether anything gets done by it.

It seems that we need large models of mental processes because we cannot keep all of the factors in our own (conscious) minds at once. But such bulky (and at times, balky) models can require a considerable segment of several lifetimes to design, construct, test, and argue about. They may be done well, but not be worth all the work that goes into them. Perhaps we should give up as ill-advised such excessively large-scaled or fine-structured systems of "bold conjectures" and instead test hypotheses of the theory in the usual empirical ways. But now another dilemma arises. Whole theories cannot be tested; only single hypotheses can have empirical content. So which hypotheses should we test and how? Empirical testing is not the knockdown arbiter of the theoretically imagined that it was once held to be, because one must pull in fragments of still other theories, subsidiary assumptions, and additional data for the testing process. "*Aut viam aut faciam.*" Either I shall find a way or make one. Which approach is best for solving our problems? Probably both, if they can become reciprocally supportive in an inquiry procedure designed so that either can call on the other - a procedure that, not so surprisingly, resembles the indirect recursions to be found in modern algorithms.

#### NOTES

1. Something has to be canned at some level. Would the AI purist allow us to can words or must letters and phonemes be the basic units?

2. As mentioned earlier, persons are in part artifacts, man modulated and man shaped.

The preceding notes were added in proof and hence not seen in the preprint circulated to the commentators; they should therefore be regarded as part of the Author's Response.

## Open Peer Commentary

*Commentaries submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article.*

### Going after PARRY

Robert P. Abelson

*Department of Psychology, Yale University, New Haven, Conn. 06520*

PARRY's delusions of persecution may have a kernel of truth. PARRY has been subject to a lot of criticism. PARRY remains elusive and hard to pin down. Somehow, PARRY has not been fully accepted in psychological or artificial intelligence circles.

I feel that PARRY is an important model because it

incorporates intentions and affects in its program. Its interpretation of input depends not only on the conversational context, but on its internal states, in turn a residue of the entire interaction. It must contend with its own competing motives, on the one hand to avoid humiliation, and on the other, to reach out for help from the doctor. These humanoid complexities are not characteristic of any other program dealing with natural language. Thus PARRY represents a beginning of a part of the fascinating and uncertain project of building a robot that manages (or mismanages) its own motives and feelings. In so doing, it challenges Searle's (1980) assertion that such a project is intrinsically doomed. The PARRY program also proposes specific mechanisms for some of the motive-handling skills needed and deserved by robots, in the opinion of philosophical optimists such as Sloman (1981).

In the long view, therefore, PARRY is salutary and ahead of its time. Its status in and of itself, as an AI model and as a psychological theory, is harder to evaluate. There are two main questions: Does it really explain paranoia? Do the mechanisms generalize to anything else?

The Turing-like test of indistinguishability is, as Colby clearly recognizes, a weak one. Part of the problem is that the psychiatrists trying to make the distinction between PARRY and real paranoids don't have a well-specified theory by which to make the judgment. If they did, Colby would not have needed to formulate the model!

Consider one of the specific low-level claims in the PARRY program, namely that when shame is high, a question such as "Are you sick?" will be interpreted as an attribution of craziness. The program will then respond with even higher levels of shame, reinforcing the paranoid cycle. I presume that the general principle behind this specific implementation is that paranoids are potentially extremely sensitive to insult, so sensitive that they fail to parse input correctly if a part of that input can be misconstrued as the declarative concept (BE I CRAZY). Thus, if the doctor said, "If anyone says you are crazy, I'll punch him in the nose," a vigilant PARRY would extract (BE I CRAZY). Whatever else was also coded from the sentence wouldn't matter much, because the productions running off (BE I CRAZY) would dominate processing.

The shame-vigilance characteristic modeled in PARRY thus makes strong claims about particular interpretative distortions. Do real paranoids behave in that way? Perhaps neither psychiatrists nor others really know, because the limits of distortion have not been systematically tested by such double-edged inputs as that suggested above. Why not try to find out? The instinct of the experimental psychologist would be to study the responses of real paranoids to systematically manipulated interviewer behaviors. If paranoids are sensitive under the same circumstances as PARRY, that would provide a stronger boost to Colby's formulation than the weak Turing test.

Of course, it is not easy to implement this experimental instinct, for reasons both of ethics and of practical difficulty in guiding conversational flow via prearranged probes. As an alternative, perhaps Colby's collected interviews with real paranoids can reveal not simply whether the psychiatrist thought the person was a computer program, but whether someone knowing PARRY would find the person PARRY-like in detail.

Whether the mechanisms in PARRY's "normal" mode are general can only be answered by trying them in programs with other purposes. Imagine a simulation of someone taking a job interview and trying to impress the interviewer. Or going to a cocktail party and arguing that this, finally, could really be the year of the Red Sox. Or questioning a doctor for the truth about the illness of a close relative. And so on. Would such other conversationalists – call them LARRY,

BARRY, HARRY, CARRIE – partake of the same affective dynamics and the same metasystem as PARRY? Naturally the content of the concepts and productions would differ, but the interesting question is how much carry over of mechanisms there would be. I believe that Colby's intention is to promote, at least in broad terms, such transfer.

## Simulation?

Joseph Agassi

*Department of Philosophy, Boston University, Boston, Mass. 02215;  
Department of Philosophy, Tel-Aviv University, Tel-Aviv 69978, Israel*

Computer simulation of human conduct and artificial intelligence experiments need no justification; they are amply rewarding even as mere games. It is possible to ascribe to them some clear-cut purposes, nevertheless, such as testing deductive validity; freeing their user from redundant or erroneous intuitions; making us aware of tacit presuppositions which may be in need of testing or which are true by virtue of mere accidental circumstances, which may at times lead to the profitable alteration of these circumstances; bringing into focus unnoticed aspects of very well known phenomena. In addition, the very theory of analogy is thereby developed, with far-reaching consequences for a wealth of fields from logic to engineering – including computer sciences, of course.

When judged as goal directed rather than as games, diverse computer simulations may be judged as more efficient or as less so. In particular, they may be judged as more or less efficient than the traditional means to given ends. There can be no general rule for comparing traditional and computer ways of doing things, and therefore it may very well be highly advisable to estimate efficiency before deciding which means to use. For example, the aim may be to test a given theory, and whether or not the computer is helpful in this task, or whether simulation is the best way to use the computer, cannot be decided a priori. At times a model may be easily tested with the aid of simple traditional means, and only after it is refined enough and made sophisticated enough, can the aid of the computer be profitably enlisted. The opposite, too, may be the case: a precise variant of a vague model may be refuted, without precluding the possibility that it may have another, unrefuted and possibly true, precise variant. All this makes the activity rich and interesting, yet it also calls for some measure of sophistication.

I find very impressive and wish I could find credible the claim of Kenneth Mark Colby that his computer simulation of a paranoid patient who follows the strategy of minimizing or forestalling shame-induced distress fooled five experienced psychiatrists, since, after interviews they had conducted with the patients for 30 to 60 minutes via teletype instruments they could not distinguish between human and computer. This is certainly an impressive achievement, much ahead of the ELIZA simulation of a Rogerian psychiatrist devised by Joseph Weizenbaum (1976) that so impressed the public. Yet too much is problematic in the present case, and Colby's lengthy article presented to me for commentary includes only two paragraphs of a report on the test itself. All I can do, then, is raise a few friendly queries, in the hope of getting satisfactory elucidations or generating useful developments in the near future.

To begin with Colby's reliance on unanimity among psychiatrists on matters diagnostic: it is very well known that this is no small matter. Many people assume that hospitalization at least proves a case uncontroversial, but this is not so, since hospitalization is more problematic than diagnosis in somatic and in mental illness. In diagnosis, only very extreme cases gain something approaching consensus among psychia-

trists: for example, childhood autism, which is clearly associated with a horror of any contact, physical or verbal, peculiar dance movements, and more; or perhaps hysterical epilepsy, which shows the whole range of symptoms of epilepsy without the brain patterns characteristic of epilepsy; or *paranoia vera*, which is a pervasive delusional *idée fixe*, logically sustained. Yet not all extreme cases win diagnostic consensus: the severest case is that of catatonia. In catatonia total freezing leads to death within a short period unless the patient is forcibly handled, yet we do not know how to distinguish the real case from the hysterical one, that is, from its simulation under certain conditions. In any case, Colby studied cases that were hardly what a psychiatrist would call extreme.

The case Colby presents is that of persecutionism. He calls it paranoia, which is an idiosyncrasy, since in the normally accepted sense paranoia is the delusional entertainment of an *idée fixe*, and Colby says nothing about this. Of course, I do not in the least mind his idiosyncrasy, but I do wonder how it affected the psychiatrists with whom he worked. This is particularly significant, since persecutionism can, but need not, be delusional or systematic (with or without being paranoid in the usual sense of the word), and Colby does not exhibit any sensitivity to the differences. In particular, whereas a person who is not persecuted but feels persecuted may be said to suffer from persecutionism, a person who is persecuted and feels persecuted will be judged as behaving normally – so much so that a person who is obviously persecuted yet shows no signs of being aware of it may be said to be suffering from *la belle indifférence*, which many consider a form of severe hysteria. Now, it is hardly conceivable that a patient in a mental home will not feel justly persecuted, especially if he is, like the one Colby describes in his Table 1, engaged in shady business.

An objection may be raised here: an expert may be trusted, one may object, to know all this and judge mentally ill only people who exhibit certain symptoms to a sufficiently high degree. This objection is valid, and may very well be correct. As it happens, it was tested – and even in Colby's own neighborhood. The result of the test was a clear-cut refutation of the objection. Perhaps the empirical data were not that good, but certainly they permit one to entertain reasonable suspicions – especially in view of the fact that Colby did not test this objection: he did not as I understand it try to pass a normal person for a persecutionist.

Finally, I am not clear about the connection between persecutionism and the shame-evasion strategy. That strategy is common, of course, in mental homes; it is also common in jails, poorhouses, brothels, orphanages, Salvation Army centers, and even on street corners. Since, however, mental patients, drunkards, and other confused people are easier to emulate than people with clear-cut character traits, as any actor knows, it may well be advisable to try to simulate their conduct first. But a clearer awareness of the task may be more useful and more interesting.

### Issues in computer modeling of cognitive phenomena: An artificial intelligence perspective

Jaime G. Carbonell

Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Penn. 15213

This commentary addresses three aspects of Kenneth M. Colby's article "Modeling a Paranoid Mind." First, issues involved in constructing and evaluating a computer simula-

tion system as an investigative tool for psychological theory formation are considered. Second, the natural language parsing component of PARRY is briefly compared with other parsing methods in artificial intelligence. Third, two alternative models of encoding belief systems are surveyed.

#### *Evaluating computer models of psychological phenomena.*

Colby opens his article by extolling the virtues of computer simulation as an investigative tool for exploring cognitive phenomena. I wish to underscore his basic thesis, outlining a somewhat different argument, and inserting a word of caution to guard against unwarranted optimism. It is clear that a performance model replicating the observable behavior typical of a cognitive phenomenon under investigation provides some insight into that phenomenon and necessarily incorporates at least aspects of a theory. But exactly what insights are provided by a working simulation model? How can one evaluate such a model or the (often implicit) underlying theory?

First, consider some positive aspects of computer simulation, which, in my opinion, far outweigh potential pitfalls. Philosophy of science often measures the adequacy of a theory in terms of its descriptive and explanatory power.<sup>1</sup> However, the power of a computer as a theory tester, that is, as a vehicle in which complex theories can be stated in procedural form, expands the horizons of theory development and evaluation beyond those defined by the philosophy of science. Computer simulation enables us to speak of the *procedural adequacy* and *fine structure* of a theory.<sup>2</sup> A procedurally adequate theory provides a constructive method of recreating a phenomenon, whereas a descriptively adequate theory need only make statements characterizing the manifestations of that phenomenon. Computer simulation enforces procedural adequacy, which in turn requires commitment to fine-structure detail, thus producing more complete, useful, and testable theories. Moreover, the computer proves to be a relentless judge of internal consistency in large-scale models.<sup>3</sup>

On the cautionary side, it is crucial to ascertain whether a computer simulation truly provides a theory of the phenomenon in question or merely replicates epiphenomenal manifestations by a process unrelated to the cognitive mechanism it purports to model. More specifically, one must ask the following analytical questions of a computer simulation system:

Does the simulation system have generative power? That is, does it produce appropriate behavior directly responsive to the input over a much wider range of situations than the input set considered by the system designer?

Does the program embody a viable constructive theory of the cognitive phenomenon, or does it merely mimic the input/output behavior? Do the intermediate processing stages correspond to psychologically plausible states?

Can the theory underlying the simulation make concrete predictions that can be experimentally verified or disproven?

Does the computer model explain the *fine structure of the cognitive phenomenon*? A computational theory is in principle procedurally adequate in that it postulates the detailed process giving rise to the phenomenon under consideration. But, as Colby notes, are the low-level procedures part of the theory, or are they merely "implementation details"?

Given the limited scope of the target article and this commentary, I cannot do justice to Colby by attempting a full analysis of PARRY based upon the above criteria. However, I will mention a few highlights. PARRY is more transparent than most other artificial intelligence simulation models. As such it lends itself to closer scrutiny – a condition that should be encouraged in AI. Colby's analysis is largely based on input/output behavior; it does not address the fine-structure issues raised above. Therefore, his cannot be considered a validated theory of paranoia at the level of detail in which the program is written, or at any level of abstraction preserving

full integrity of the processes encoded in PARRY. However, PARRY does meet the generative test, in that it responds to unforeseen input in a constructive and reasonable (if somewhat single-minded) manner.

**A note on parsing natural language.** Since Colby devotes a substantial portion of his article to a discussion of his language analysis mechanisms, some mention of other parsing systems in AI ought to be made, in order to set PARRY's parser in perspective. As Colby correctly states, PARRY's pattern-based parser is meant to be an engineering tool in the service of a working model, whose purpose is to simulate a different aspect of human cognition. There are, however, other efforts in AI whose objective is to simulate human language understanding, such as Riesbeck and Schank's (1976) expectation-based semantic analysis system, which is based on Schank's (1975) conceptual-dependency model of meaning representation.

Other approaches taken to automated language analysis include more syntax-based linguistic approaches, such as augmented-transition networks (Woods, Kaplan & Nash-Webber 1972) and task-oriented semantic-grammar approaches (Hendrix, Sacerdoti & Slocum 1976). Presently, a major thrust in computational linguistics is the creation of robust flexible parsers, capable of formulating meaning representations from ungrammatical input (Carbonell & Hayes 1981). It is interesting to note that Colby sought to process input that did not necessarily conform to a prespecified grammar, as long as some meaning could be extracted. The more sophisticated language-analysis systems today share at least this important objective with Colby's earlier work. [See also Arbib & Caplan: "Neurolinguistics Must Be Computational" *BBS* 2(3) 1979.]

**Alternative methods of encoding belief.** Belief-system representation has been an active research area in artificial intelligence. Colby largely pioneered the effort with his initial model of neurotic belief. However, his later efforts do not appear to have centered on developing a general model of human belief systems. It is perhaps worth mentioning two other models of belief based on different computational mechanisms. Robert Abelson (1973) developed a system called the "ideology machine," whose purpose was to model Senator Barry Goldwater's political pronouncements in response to statements about communist actions in a cold-war scenario. Abelson's objective was to model "hot cognition," in the sense that internal conviction and affect played major roles in formulating responses. However, his central contribution was the *master script*, a data structure that encapsulated the essence of ultraconservative political belief toward a perceived communist threat. That ideological structure could then be applied to a large number of relatively novel situations.

In the development of POLITICS (Carbonell 1981), I have taken a different approach toward modeling belief. A set of *goal priorities* captures the essential differences among different ideologies. An ideologue attributes a set of motivations (and priority rankings among the motivations) to various political actors, such as the United States and the Soviet Union. Assuming that all behavior is goal directed, and that more important goals dominate less important ones when goals conflict, every political situation is interpreted by each ideologue when he is able to attribute known goals to the actions of the political actors. Since each ideologue perceives goal priorities differently, interpretations of political actions may differ radically. It is significant that differences in goal priorities – with no differences in inference procedures – were sufficient to enable the POLITICS program to model left- and right-wing American politicians responding to a large range of political scenarios.

There is no direct way to contrast PARRY with POLITICS or the ideology machine, as the aims of the investigators were

far from identical. However, it should be noted that three distinct computational mechanisms proved capable of simulating belief-guided behavior, and none of the three can be proven "correct" in the sense that one subsumes or invalidates the other two.

#### NOTES

1. Chomsky (1965), for instance, eloquently applies these notions in defense of his transformational theory of linguistics.

2. Aaron Sloman (1981) argues that fine structure is a meaningful measure of theoretical adequacy absent in Kuhn's (1977) treatise.

3. The reader is referred to Meehan's (1976) TALESPIN and Carbonell's (1979) POLITICS for graphic examples of the utility of computer simulation as a method to test the internal consistency and validity of alternative versions of computer models partially embodying cognitive theories. Experience has shown that there is no substitute for an automaton that carries theoretical assertions to their logical, often unforeseen conclusions.

### Paranoia concerning program-resistant aspects of the mind – and let's drop rocks on Turing's toes again

Keith Gunderson

*Department of Philosophy and Minnesota Center for the Philosophy of Science, University of Minnesota, Minneapolis, Minn. 55455.*

Nearly twenty years ago, in a parody of A. M. Turing's (1950) test for intelligence in his "Computing Machinery and Intelligence," I concocted a parallel test called "the toe-stepping game" for helping us decide whether rocks could ever imitate, and claimed

that in less than fifty years' time it will be possible to set up elaborately constructed rock boxes, with large rock-storage capacities, so that they will play the toe-stepping game so well that the average person who would get his toe stepped on would not have more than a 70 per cent chance of making the right identification after about five minutes of toe-stepping.

(Gunderson 1971, p. 44; see also Turing 1950; Gunderson 1964) I stick by that prediction and anticipate a successful rock box design in less than thirty years. The question at hand is whether Colby's already extant model for a paranoid mind is more or less exciting for psychological theory than such a box would be.

The underdeveloped distinction between what I've called *program-receptive* and *program-resistant* aspects of the mind was intended as a first stab clarification in cybernetic terms of what then seemed, and still seems to me, to be an interesting and important difference between features of the mind that are behaviorally "saturated," such as various kinds of problem solving (theorem proving, chess playing, and the like) and features that seem more riddled by conscious experience or "qualia" (non-problem-solving nonbehaviors such as *having* a blue afterimage, or *feeling* anxious, or *being* in pain; see Gunderson 1971). As noticeably imperfect as that distinction was and is, either it or some comparably cleaving complication in the taxonomy of mentality is called for, unless, as seems heartily implausible, radical behaviorism or some (now) equally unlikable sibling sidekick to it turns out to charm us all (see Dennett 1978; Holly 1975.)

But to argue that a feature of the mind – an afterimage, a pain, or an anxiety – is program-resistant, is not thereby to claim that it is model resistant *simpliciter*. What I argued in *Mentality and Machines* was that if one is to attempt to model the having of an afterimage or the feeling of an

anxiety, one will need to do so in some way different from the way in which attempts have been made to simulate theorem-proving or chess-playing behaviors and the like. My suggestion was that the modeling or simulation problem for program-resistant aspects of the mind was more like the problem of somehow imparting to the machine analogues of basic capacities presupposed by possible software strategies, than it was a problem of writing routines or programs that use such capacities. If there were, to begin with, pain receptors or afterimage generators or anxiety inducers, or some illuminating analogue thereof, then, but only then, would it make sense to attempt to simulate the behaviors they typically give rise to. (Whether or not Colby's model includes illuminating analogues of affects is one of the important issues.) Compare: only if there's already the capacity for storing bottles, cans, or cups of and for Coca Cola in a machine, and a slot for the coins, will it make sense to proceed to endow it with its robotic cromagnon competencies for dispensing drinks once the requisite legal tender tinkles down the chute. The methodological howler I accused some computer simulation psychologists of making was that of trying to employ almost exactly the same tactics for simulating program-resistant aspects of the mental that had been used with varying degrees of success for program-receptive aspects of the mind (Gunderson 1971). I did not claim that there were any aspects of mind, including program-resistant ones, that were in principle unmodelable or unmechanizable. And I certainly never meant to suggest that for any model of the mind to be an interesting one, it would have to replicate the mind in some degree or other. Ironically, Colby seems to have much more affection for a replication ideal for modeling the mental than I've ever had, and invokes it at the level of input/output behavior patterns. His use of Turing-like tests is his methodological means by which to tell if that ideal has been met.

Whatever any of the many things successful modeling of the mind might mean, I do not think it involves Turing's test or any variation thereon. To invoke that test as a criterion for success in computer-simulation modeling obliterates what I still see as a substantive distinction between the goals of artificial intelligence (AI) and the goals of computer simulation of cognitive processes. The initial goals of AI were to devise machines capable of performing various tasks which hitherto required human intelligence to perform, whether or not the machines mirrored human cognitive processes. Thus success at AI might obtain without it having any clout whatsoever for psychological theory. CS (computer simulation) on the other hand, as practiced by Simon, Shaw, and Newell, et al., was avowedly more ambitious with respect to mirroring human psychological processes. And it was the failure to appreciate this that vitiated Wang's (1960, as cited by Minsky 1963) early criticisms of Simon, Shaw, and Newell's work on the "Logic Theorist" project. (See Minsky 1963; Gunderson 1971.) Colby, so far as I can tell, happily reiterates the conflation of the goals of AI with the goals of CS. But, then, he's hardly alone in this, for any form of reliance on Turing's test is likely to push one in that direction, since passing Turing's test just amounts to obtaining the psychologically tepid goal of AI. [See also Pylyshyn: "Computational Models and Empirical Constraints" *BBS* 1(1) 1978.]

Two rather different but complementary objections to that test should be made indelible. The first is simply that input-output similarities between machines and cognitive-affective beings are neither necessary nor sufficient for establishing interesting internal process, state, or event correlations. And it is these internal goings-on, I still believe, (contra Ryle et al.) to which our mentalistic vocabulary somehow mysteriously refers and applies (see Dennett 1969; 1972; Gunderson 1972; Ryle 1949). To be content with Turing's test criteria for assessing modeling success in CS is rather like thinking that slapping two halves of some analogical bun together makes a

good model of a hamburger. ("Hey, where's the meat? the onion? Also, I wanted a pickle!") Conversely, to be discontent with a simulation model of cognitive processes because it failed to pass Turing's test would be a mistake. For there might well be an illuminating correspondence between internal processes, states, and events of some computer model and human processes, states, and events, yet deficient mapping at input and output levels. A cooked hamburger patty with onion and pickle upon it might give us a pretty good grasp of a hamburger even if an analogical bun were lacking and the whole repast were lodged within pita bread. ("Hey, this tastes pretty much like a Big Mac!")

The other objection to Turing's test is this. If there are program-resistant aspects of the mind, the criteria to be met in modeling these cannot be found along behavioral problem-solving lines. Turing's test and all the modifications of it that I'm aware of have with a vengeance been along just such lines. To model the symptoms or effects of our consciousness or "qualia," paranoid or other, is not to model consciousness or "qualia." To model a pain in one's toe, it is not sufficient to model the tokening of some sentence such as "Ouch! my toe hurts" (see Descartes, 1637). And to insist on this is not the same as to insist that a model replicate what it models.

To have said all this, however, is not, unfortunately, to have contributed much to an understanding of what a model of consciousness or "qualia" might be like. Sometimes I'm inclined to think such understanding can never be forthcoming, and that the very notion of modeling is hopelessly harnessed to third-person perspectives, whereas consciousness or "qualia" are essentially first-person ones. In other words, to be able to sketch what a model including consciousness would be like, is tantamount to being able to sketch a model of human (or bat) subjectivity (see Gunderson 1970; 1972; 1974; Nagel 1974; 1979). And this we do not know how to do, which may be, in the last analysis, the same sort of ignorance as our ignorance of the nature of the mind-body relationship itself. As Thomas Nagel (1979, p. 177) puts it:

The idea of how a mental and physical term might refer to the same thing is lacking, and the usual analogies with theoretical identification in other fields fail to supply it. They fail because if we construe the reference of mental terms to physical events on the usual model, we either get a reappearance of separate subjective events as the effects through which mental reference to physical events is secured, or else we get a false account of how mental terms refer (for example, a causal behaviorist one).

The central problem seems to sum up like this: to aim at (now, and maybe always) a replication of program-resistant aspects of mind, or consciousness, in a model is hopelessly utopian. But all extant alternative ways of incorporating it seem hopelessly prosaic. They all have that representation-by-fiat-only ring to them: as if by simply naming some operation or item in a program or flow chart and saying "This (affect, etc.) is represented in the program by *that*" and then proceeding merrily to describe input-output patterns, one believes that an illuminating model of affect has taken place. This is all too much like drawing a pineal gland, making a dot in it, and saying "This," pointing to the dot, "represents the human soul as it is joined to the human body." Well, of course it does, but that tells us nothing illuminating about the human mind, but rather something about how easygoing and unilluminating certain forms of representation can be. Herbert Simon's (1969) representations of motive and emotion in his "Motivational and Emotional Controls of Cognition" seemed to be unilluminating in just this way. And Colby's "conceptual representation" of feelings of shame seem to me to fare no better (see Gunderson 1971).

In spite of the heavily critical emphasis in the foregoing, nothing I have said really touches upon whether Colby has or hasn't made an important contribution to the *taxonomy* of paranoia. Because he might have, and because CS models in

general may, perhaps, be usefully viewed as at least in part a new way of doing taxonomy of the mind, we can be thankful that it is not necessary to solve the mind-body problem before proceeding with CS research programs. Not that anyone now engaged in them would be likely to be halted in his tracks were I to argue otherwise.

### Colby's paranoia model: An old theory in a new frame?

C. E. Izard and F. A. Masterson

Department of Psychology, University of Delaware, Newark, Del. 19711

A significant portion of Colby's paper is devoted to the author's views about model building in general and AI model building in particular. We are sympathetic with these views and agree that AI models can make useful contributions to the understanding of psychological processes. AI modeling focuses attention on details that the theorist might otherwise be tempted to ignore (as Colby notes, some of these "details," once squarely confronted, turn out to be extremely complex.) Another contribution of AI is an emphasis on *synthesis*. Psychologists analyze phenomena into parts, but often fail to specify how the parts can be put together to synthesize whole phenomena. Because of the emphasis on performance, the AI modeler must specify not just the detailed operations of the parts but also the detailed manner in which the parts of the model interact to produce intelligent behavior. The most important contribution of AI to psychology may be that it restores a balance between analysis and synthesis.

Of the many AI formalisms, that of production systems seems especially relevant to scientists in the behavioral and neural sciences. A production system can be viewed as a modernized version of stimulus-response rules. The improvement lies in the fact that the entities on either side of the hyphen can be identified with internal or mental events, in contrast to the peripheral events of the old-fashioned stimulus-response approach. The left side of a production rule describes possible contents of working memory. The right side can either add new contents to the working memory or initiate an action. Thus, productions are cognitive analogues of stimulus-response associations, and might therefore appeal to behavioral scientists who appreciated the simplicity of the latter approach but have rejected its peripheral bias. The potential appeal of the production-system approach to brain scientists, as well as to behavioral scientists, lies in its emphasis on parallel processing by a large number of simple modules. Each production encodes a small amount of procedural knowledge. In principle, vast numbers of productions can scan working memory simultaneously.

A major difficulty for the AI approach is that the models tend to be complicated and therefore difficult to describe. (Perhaps this is a price to be paid for models that are sufficient to emulate some of the details of intelligent behavior.) Early in his article, Colby declares that one should be able to communicate the assumptions of an AI model in English. However, this is more easily said than done, and Colby later apologizes that his "piecemeal sketch" fails to convey much about how his model operates. Unfortunately, this is true, and it is therefore difficult to evaluate the psychological assumptions underlying the model.

Given the rather brief description of the model, the reader's judgment must rest on the results of the Turing tests. For this reason, and also because this appears to be one of the first serious applications of Turing's proposed test to an AI model, it is a pity that these tests are not described more fully. Colby says that he has performed many such tests, yet he reports only one of them, and does not describe that one in much

detail. What happened in the other Turing tests? And what are some of the controversies that surround the idea of a Turing test? Colby says that he has grown tired of these controversies, yet he never describes them. Given the importance of Turing's test for judging the sufficiency of AI models, a description of these controversies, or a reference to such a description, should have been included.

Although the model succeeded in the "Turing-like tests," we have some reservations about the magnitude of this feat. We believe that the paranoid mind may be the easiest one to simulate, because of certain rather rigid affective-cognitive orientations that characterize paranoia. Thus, it should be relatively easy to develop a script that depicts the paranoid. The potential importance of Colby's accomplishment is that he has constructed a model in which affects mediate paranoid responses to relatively neutral input.

We think Colby's inclusion of affects or emotions as the source of motivation for the model's behavior is a step in the right direction. It is a timely step too. The contemporary current of belief in the cognitive sciences is shifting toward the view that information-processing models without emotion variables have limited utility (Izard 1977, note 1; Norman 1980; Zajonc 1980).

We have some questions regarding Colby's explanatory theory of paranoia, whose illustration was his "primary purpose." The affects programmed in the model are fear, anger, shame, distress, interest, and enjoyment. As in the more general theories of Tomkins (1962; 1963) and Izard (1971; 1977), each of these emotions is treated as having unique motivational properties. However, one of the emotions in Tomkins's (1963) cogent theory of paranoia is missing in Colby's model—contempt-disgust. (Tomkins's theory is not substantially affected by the fact that Izard 1971; 1972 later demonstrated that contempt and disgust are separate, though frequently associated emotions.)

Tomkins's theory, which I find more convincing, not only contains an additional affect, it differs in principle from Colby's. Colby's holds that the paranoid's thought and behavior consist of "strategies for minimizing and forestalling shame-induced distress." In contrast, Tomkins maintains that in the paranoid the emotion fear-terror is conjoined with shame-humiliation. The paranoid's affective experiences have been organized into a "monopolistic affect theory" that dominates consciousness. The paranoid engages in "unrelenting warfare in which the individual generates and tests every conceivable strategy to avoid and escape total defeat at the hands of the humiliating bad object" (Tomkins 1963, p. 455). For Tomkins, the source of shame-humiliation is mainly perceived as the *contempt* of others, as well as *self-contempt*.

Thus the paranoid is fundamentally one who perceives the self as constantly teetering on the brink of inadequacy, insecurity, and total incompetence. The feeling of ineptitude reflects the shame, and the insecurity and the threat of total loss of self-control reflect the fear.

Insofar as the Turing-like tests were effective, we have to attribute some success to Colby's model. But on closer examination, Colby's and Tomkins's explanatory theories of paranoia have one important principle in common—the motivational impact of shame on cognition and action. In Colby's theory the paranoid seeks to reduce distress, but it is shame-linked and *shame-induced* distress. Therefore, in both theories shame anticipation and shame avoidance would be common strategies. Tomkins's concept of contempt-induced shame has some overlap with Colby's notion of the paranoid's experiencing shame as a function of the "verification" of self-defectiveness beliefs. Furthermore, Colby's concept of distress is vague, defined in one place as a function of shame or "an indication of the 'painfulness' of the shame-humiliation affect." Therefore, for Colby, distress is nothing

more than intense shame-humiliation, which can indeed be very painful. If our interpretation and alignment of concepts and principles are correct, Colby's underlying explanatory theory paraphrases that of Tomkins, who should be credited with the original.

There are several specific points in the paper that need clarification. If distress is equivalent to intense shame-humiliation, it does not follow that it would be activated by the "summation of values of three negative affects" (shame, anger, fear).

Colby should document his claim for the relative importance of verbal over nonverbal cues and recognize that the importance of the latter may vary greatly with age and diagnosis.

It is not really clear whether and how the model allows for emotion-induced emotion, which is important in Tomkins's theory. Colby suggests that this process does occur in his model (see "The problem of model equivalence").

In fine, we found the paper an interesting and challenging demonstration of AI theory at work, but thought it unduly wordy. Most important, the scholarship, and probably the model, would be improved by more explicit handling of the relation between Colby's and Tomkins's theories of the paranoid mind.

#### NOTES

1. Izard, C. E. The primacy of emotion in human development. Paper presented at the meetings of the Society for Research in Child Development, April 1981.

### On the generality of PARRY, Colby's paranoia model

Manfred Kochen

*Mental Health Research Institute, University of Michigan, Ann Arbor, Mich. 48109*

Artificial intelligence is primarily an engineering art. The most impressive aspect of PARRY is its well-engineered parser, which accepts a large variety of English inputs about delusions of persecution by the Mafia. The SUMEX-AIM project, of which PARRY is a part, prides itself in the claim that artificial intelligence (AI) has progressed to the stage where it is useful. We are indeed witnessing the introduction of intelligent switching centers in telephone networks and intelligent software for scheduling many simultaneously run programs in computers, as well as intelligent terminals and robots. The practicality of computer-assisted diagnoses based on expert systems in ophthalmology and internal medicine is being demonstrated by engineers.

Since our earliest computer simulations of cognitive tasks (Kochen 1960a; Kochen & Galanter 1958), there has been a clear distinction between the simulation of human behavior and the behavior of a computer in solving a problem that is compared with the behavior of a person solving the same problem. The problem faced by PARRY is "how to behave in a psychiatric interview," both in a normal and in a paranoid mode. (It is not clear whether that is interpreted normatively or descriptively, and it is not a well-defined problem.) Even if trained interviewers fail to discriminate at the input-output level between the behavior of the model and that of a person, the claim that Colby has created a hypothetical "patient" must be challenged. If "patient" is interpreted in its ordinary, nontechnical sense as a "person who is ill or is undergoing treatment for a disease," then the claim is untenable, because PARRY's behavior does not resemble that of an entire person, even superficially. In the description of this model, "hypothetical patient" must have been intended as a technical term.

The mildest criticism of Colby's article, then, is its uncritical use of technical terms and its failure to distinguish between technical and ordinary terms. Where PARRY behaves as expected, it is because it has a well-structured data base, a special internal structure appropriate to delusions of persecution by the Mafia, a reliable means of dealing with a restricted subset of English, and an adequate inference procedure.

That the model cannot behave appropriately outside of its restricted domain of discourse can easily be verified by interacting with PARRY 3 on SUMEX-AIM. Here are some sample dialogue lines:

I: Hello, I am Dr. Chen. I just came from China. Would like help you.

M: I'm neutral about you.

I: Why did you say that?

M: You haven't said much about yourself.

I: I was born in Shanghai. Educated in Peiping. What else would you like to know?

M: That's not true. I don't smoke. (Then goes into standard line about bookies and Mafia.)

Another sample follows:

I: Would you rather be outside the hospital?

M: Yes I am. This place is not for me.

I: Where would you live?

M: I live in San Jose.

I: How would you support yourself? When you are out, will you be able to pay rent and buy food?

M: It's about eight o'clock.

I: What does that have to do with my question?

M: It's not helping me to be here.

I: What would help you?

M: I don't really need help like they think.

I: What do they think?

M: Do you know anything about bookies? etc.

The nonsequiturs in these dialogues reflect the restricted domain of discourse rather than a deliberately designed paranoid or other mentally abnormal reaction. After two or three encounters with PARRY, it is easy to predict its highly stereotyped and inappropriate responses, such as "I don't think I should have come here," "I have been somewhat upset," "I feel good right now," "I get afraid once in a while," "Let's not have another encounter session."

There is little the interviewer can say to probe into the connection between this delusion about being persecuted by the Mafia and the strategy for minimizing and forestalling shame-induced distress that presumably maintains the delusion. Indeed, the interviewer cannot ascertain that it is a delusion. Nor can he try anything that might be useful in planning treatment or making a prognosis.

The ability to generalize, classify, or abstract seems to me fundamental for any model of mind, normal or deviant. How to incorporate this into models has been understood to some extent since at least 1960. It might be incorporated into a future version of PARRY (Kochen 1960a; 1962; 1974). Operational programs that form, revise, and change the weights of hypotheses about their environment have been constructed; they demonstrate a moderate degree of generalizing ability. Beliefs about the interviewer's intentions are special kinds of hypotheses. PARRY could be modified to enable it to change the strength of such beliefs as "I am worthless," or "He thinks I'm dishonest." This, in turn, could raise the level of shame, which could temporarily strengthen some beliefs. Certain mechanisms control the dynamics and determine whether the vicious circle continues to escalate or whether it snaps back to normal. These control mechanisms may be the essence of the problem.

If the delusion of persecution by the Mafia is but a symptom of M's procedure for forestalling or coping with

shame-induced distress, and if that delusion somehow became untenable, we might expect another delusion to substitute for it. M might, for example, switch to delusions of persecution by the KGB or the wrath of God.

Colby asserts that it is trivial to add volumes of facts to the model-patient. Indeed it is, but not much more so than having put in the facts that the model uses now. There is more to a paranoid patient than the beliefs about being crazy, dishonest, worthless, or stupid that are stored in a computer and the facts introduced by the programmer to confuse an interviewer about whether he is dealing with a human paranoiac or PARRY. What is missing and essential for substantiating the claim that Colby has created a "patient," is the model's ability to generate and modify its own beliefs. To be sure, the designer would still embody in the model's generating procedures his theory of paranoia as part of a more general theory of mental function. Computer programs such as AMNIPS (Kochen 1962), ASP (Hantler & Kochen 1973), and SEQUENTIAL (Kochen 1971) embody this generative capacity to some extent. Moreover, they provide for a limited transfer of appropriate behavior from one domain of discourse to another. That is most important. The volume of world knowledge input that must be provided in the form of facts to make possible even the simplest realistic simulation is by no means trivial (Kochen & Stark 1978). It is the generality of the model – its applicability to more than one manifestation or example of paranoid behavior – that would be a minimal requirement for a model of a paranoid mind.

### How smart must you be to be crazy?

Robert Lindsay

*Mental Health Research Institute, University of Michigan, Ann Arbor, Mich. 48109*

Inevitably, Colby's target article will raise more questions about method than about the specific model of paranoia presented. Colby anticipates most of these questions and generally addresses them ably. I believe that it is appropriate in considering such issues to apply the same standards of theoretical felicity to programmed models as we do to mathematical or literary models. Programming languages offer much broader expressive power than mathematical languages, and much greater inference-making capabilities than natural languages (yet unfortunately they lack the perspicuity of either). Even so, the problems of theoretical psychology remain nearly intractable, and it is pointless to ask programmed models to bear additional burdens not expected of other theoretical media.

One persistent objection is that no computer can be truly human; therefore no computer can be paranoid. Colby sidesteps this objection by not claiming that his program is either paranoid or human. I think this is a sound position, for we do not raise similar objections against mathematical or literary theories. (No one claims that Maxwell's equations "are" radiation.) Therefore Searle's (1980) assertion that no computer can "have" intentions is moot since no claim is made that PARRY "has" intentions – or "has" paranoia, for that matter. The claim is merely that PARRY exhibits (models) paranoidlike verbal behavior. Of course, those who feel that no scientific light can ever be shed on the dark mysteries of the mind must simply be disenfranchised from this debate.

Still, while many readers will grant that it is legitimate to model concepts without claims of creation, they will yet argue that paranoia is of necessity a property of a total personality, and cannot be abstracted from such an embodiment and studied separately. This objection is not so easily dealt with, and analogous reservations arise with regard to

other programmed psychological models. Colby wishes to deal with this claim by denying it, and we had better hope he is right if anyone, by any theoretical device, is going to be able to make progress in understanding paranoia or any other complex trait. But whether PARRY does indeed capture a sufficiently complex part of personality to shed light on paranoid thought processes is a separate question that can only be answered when some light is actually shed, or extensive attempts have failed.

In the context of a mathematical or literary theory, we tend readily to grant the possibility that a partial model can be a useful model, but "talking" to a program strongly suggests that the programmer-theorist must be claiming more of his model than the mathematician-theorist or the natural language theorist is. Many of us find it impossible to reject the challenge of this inferred claim when confronted with a talking process model that all but demands to be anthropomorphized in ways that an equation, even a satisfied one, does not.

In the case at hand, Colby is put in the position of needing a conversational model by the very nature of his problem: the model must be interviewable to pass the very first test of plausibility, since paranoids are identified largely by their verbal behavior. Unfortunately, the fact that PARRY was not distinguished from a real patient by psychiatrists in the interview experiment may cloud the issue by turning our attention from the model itself to figuring out why those psychiatrists were so gullible.

Nonetheless, it should be made clear that PARRY's success in the Turing-like test described does not demonstrate human conversational ability. Although the author is justified in claiming that a more stringent test was simply impossible, it does not follow that the test used should be accorded much significance. I feel that this objection should not be classified as a "mumble." Although admittedly vague, Turing's original suggestion was a much more demanding test than the one reported here, primarily because it pitted a computer interviewee and a human interviewee "head" to head, while a second human, the interviewer, attempted to distinguish them. An important feature of this test is that the computer was to attempt deception, *but the human interviewee was to cooperate to defeat the deception*, and the interviewer *was to know these ground rules*. To my knowledge, such a test has not yet been attempted, and I doubt that any program is ready to accept the challenge, certainly not with a sophisticated interviewer. These differences make all the difference between Turing's proposal and the PARRY experiment. A psychiatrist, probably unsophisticated about artificial intelligence, when confronted with a hostile, delusional interviewee who may be a mental patient, has little to go on, and certainly cannot probe deeply, risking serious offense to a human patient if it is one. Thus, while Colby does not need to make strong claims about the intelligence or completeness of PARRY's personality, it must be recognized that PARRY has indeed a very limited intellect.

The PARRY project possesses a flaw that is legion in behavioral science, namely, it lacks a clear success rule that would tell us when something has been accomplished, discovered, or demonstrated. Lacking a success rule we are hard put to say whether the project has succeeded or failed; we can only say that it appears forever unfinished. The usual success rule in science requires what John Platt (1964) calls strong inference, achieved by pitting incompatible models against one another in empirical joust. PARRY does not do this, but simply provides "evidence" that a single model is not patently false, a weak result that denies us science's only type of certain factual claim: that such and such is *not* the case.

To summarize, I feel that the methodology is sound, and that PARRY has passed its first test, though not a severe one. So far so good. But there is little more that can be said. There

have been no deep insights, long chains of inference, startling predictions, or technological applications of this work as yet. Perhaps that is asking too much. Still, one must ask that PARRY do more than it has. The interview method could provide a vehicle, not for showing the program's deceptive powers, but for discovering specific features of PARRY's behavior that fall short of human quality, and fine tuning the model to correct these. Until something like this is done, or a significant unexpected or deep result is forthcoming, descriptions of PARRY and these experiments must be viewed as pages from Colby's laboratory notebook, useful as progress reports and stimulating to our critical faculties in the methodological issues they raise, but lacking in conclusiveness and leaving open the question of how much intellect and affect are needed as a substrate for psychosis. The game is worth playing, but the ball is still in Colby's court.

### Colby's model for paranoia: It's made well, but what is it?

Peter A. Magaro and Harvey G. Shulman

*Department of Psychology, Ohio State University, Columbus, Ohio 43210*

Theory is an organized body of information such that the explanation includes not only the relationship between variables but also the relationship between theoretical constructs. Higher order constructs define themselves not only by what they are but by what they are not, that is, by being differentiated from other constructs. From the relationship between constructs, a theory is posited that has a specific type of structure, in some cases one that is defined mathematically.

From this observation, Colby justifies the development of a computer simulation of intelligence as a model for paranoid thought. Colby does not feel that the traditional empirical approach in psychopathology has provided enough valid principles. Although we tend to support his concern (Magaro 1980), it is not clear that Colby's target article demonstrates the superiority of simulation to the traditional deductive approach.

Our major concerns are with the scope of the underlying model, with its necessity, and with the nature of the evidence presented. With regard to the algorithms used to simulate Colby's theory we have little to say. Parkison's (1980) parser seems impressive but, as the author states, it is not a part of the theory of paranoia. Faught's (1978) interpretation-action module provides the algorithms that implement the theory. We do not fault these algorithms, but rather the scope of their permissible input and output and the identification of shame-humiliation as the driving force in paranoid behavior. We also wonder whether the choice of the psychiatric interview as the scenario for simulation may not place less than obvious constraints on the model's behavior.

First things first, though: let us consider what Colby means by paranoid. The term has multiple meanings. In psychiatry, a great deal of thought has gone into devising some system of classification that could encompass the varieties of the process. For example the current DSM-III (1980) lists a number of paranoid disorders: Paranoid Disorder, Paranoia, Shared Paranoid Disorder, Paranoid Personality Disorder, and so on. The author prefers to examine a mode of thought that he considers not to be schizophrenic but characterized by delusions and, more specifically, persecutory delusions. Unfortunately, it is not clear what that means in terms of paranoia. Not all paranoids have clear delusions, and furthermore, it is quite controversial to consider persecutory delusions as a sign of paranoia or paranoid schizophrenia (Kendler 1980). In effect, Colby considers a paranoid as someone with delusions of persecution and in a high degree of stress. The

selection of this particular symptom set as the corpus to be simulated places an upper bound on the validity of the simulation. That upper bound is the diagnosticity of the persecutory delusion for paranoia. Of course, the same is true for any other single-symptom behavioral complex that might be simulated. Valid diagnoses are seldom made, however, on the basis of single symptoms. Likewise, the theory on which this stimulation is built is not well validated by the simulation, however adequate, of a single symptom.

While we are therefore unclear about the actual type of paranoid shown by PARRY, or even whether he is paranoid in any traditional sense, we might agree that he suffers from persecutory delusions. The next question concerns the actual simulation. The underlying mechanism, in Colby's theory, is the interaction of shame-humiliation with the semantics of the interviewer's comments and questions. Here there is both strength and weakness in the author's simulation. The strength is in the attempt to incorporate the interaction between the emotional and the rational. As psychologists, we applaud this as an advance in simulation models. The weakness is in the identification of shame-humiliation as the sole emotional source of the persecutory delusion, much less paranoia. It seems to us that a single source is unlikely. However, it is not hard to see that the simulation could be modified to permit other emotions to have effects that are additive with shame-humiliation in order to rectify this problem. More damaging to the theory is the possibility that the same simulated behavior could be produced by changing a few links so that some emotion other than shame-humiliation (fear? distress?) simulates the paranoid mode. This is the question of necessity. Even if we accept the persecutory delusion as a valid target behavior, the current simulation need not lead us to accept the underlying theory, if other theories do as well. In the present case we cannot see that PARRY would behave differently were we to reconnect his emotional states. This is at the heart of things because the author's purpose in simulating the psychiatric interview is ultimately to validate his shame-humiliation hypothesis. Unfortunately, we are not only uncertain that PARRY is paranoid, but we also cannot determine whether he is ashamed, afraid, or angry.

A final observation on the underlying theory. In PARRY, the linkage between the critical emotion and the behavior characterizing the persecutory delusion is represented by links and pointers whose activation depends on the status of several emotional variables. The variability in behavior to a given input, provided by the interaction of emotional states in selecting which pathway to follow, is one of Parry's strong points. However, the mechanism by which this occurs is not explained in any psychological sense, though it is not hard to understand in terms of computing. This is a shortcoming of Colby's theory, though not one for which current psychological research can necessarily provide relief. Like the mechanism that selects pathological behavior, the processes that attenuate it are also given short shrift by Colby's model. PARRY's pathology can be diminished only by the passage of time, coupled with the absence of further provocation. Can this mean that the interviewer's recommended course, once paranoia is elicited, is to withdraw for a few hours? The model would be strengthened by the inclusion of more positive mechanisms for diminishing affect, based, for example, on principles of reinforcement.

One final pair of concerns. First, by restricting PARRY's permissible scripts to one (the psychiatric interview), validity is also limited. Would PARRY exhibit pathology in other scenarios, as a paranoid certainly would? Second, the interaction with a psychiatric interviewer whose purpose is to examine paranoia may be a unique determinant of the model's behavior. We cannot be sure that PARRY would simulate pathology in an interaction with a layman. The

combination of a model with a highly constricted belief system and an interviewer with a narrowly focused purpose congruent with the model's belief system may have constrained the simulation to a narrow range of possible outcomes.

The reservations we have expressed center on the implications of Colby's simulation for an understanding of the principle's underlying paranoia. The proof of the pudding is in the validation of the shame-guilt hypothesis. Our contention is that the proof has not been offered. On a positive note we recognize that PARRY is an application of a general model of cognition to one corpus of behavior. As a general model of cognition, we find PARRY interesting, powerful, and deserving of further application.

### Testing the components of a computer model

Brendan A. Maher

*Department of Psychology and Social Relations, Harvard University, Cambridge, Mass. 02138*

Colby's exercise tells us a number of things. It tells us that it is possible to write a program such that the utterances of a computer guided by it cannot readily be distinguished from the utterances of a real human being facing the same kind of questions from a clinician. It also tells us that a psychiatrist is likely to judge that these utterances come from a paranoid patient when their pattern includes the combination of a false belief, hostility, and self-reference in a specific reactive sequence.

The subjects of this exercise are the psychiatrists. The results indicate that the psychiatrists involved employed a definition of paranoia consistent with one that the investigator hypothesized that they did. Thus the reported results indicate that the sample of psychiatrists in this case behaved as they had been trained to behave, namely to apply the term paranoia to a particular pattern of symptoms. One problem with the study is the question of whether or not this tells us very much about the processes responsible for paranoid behavior in a patient. We may sharpen the question a little by noting the controversial study by Rosenhan (1973) in which he reported that it was possible to mislead psychiatrists into giving a diagnosis of schizophrenia by pretending to have certain symptoms in an admission interview. Ignoring for the time being the many questions that this study has raised we must note that Rosenhan was not suggesting that the nature of schizophrenia is properly represented by a model based upon components of acting and lying, nor does the fact that many diseases may be represented by malingering confirm that the "program" for malingering and that for the disease are essentially the same.

Colby presents his procedure as a test of a theory of paranoia. His theory asserts that in paranoia there is a "causal chain of strategies for dealing with distress induced by the affect of shame-humiliation. A strategy of blaming others functions to repudiate the belief that the self is to blame for an inadequacy." This theory, he states, is implemented in the computer model. The theory is not causal in the etiological sense, as Colby makes clear, but attempts instead to spell out what sequences of interpersonal interaction will occur when the patient is faced with a threat to self-esteem. The decision by actual psychiatrists that a person or computer showing such sequences should be called paranoid is the test of the theory.

What we cannot tell from this investigation is which of the various components (false belief, hostile response to perceived threat, and so on) is essential to the psychiatric judgment.

Would a computer programmed only to utter an implausible belief about the Mafia, but lacking programmatic instructions to express hostility or to exhibit suspiciousness in other ways, fail to elicit a diagnosis of paranoia? Or would hostile blaming of the interviewer without accompanying false beliefs also fail to generate the diagnosis? Or, more elaborately, would a program expressing false beliefs and hostility – the latter not specifically to perceived threat, but generally, and independently of the interviewer's comments – perhaps generate such a diagnosis?

Colby's theory demands that the judgment of paranoia be elicited only by the combination of behaviors occurring in a specific way. To test the theory at this level surely requires that none of the single components of the syndrome or any other combination of them lead to the diagnosis. The logic of the design is akin to that of a police lineup in which the witness is presented with one suspect only and asked to make a yes or no identification, rather than one in which the identification is made from a row of suspects, all with some overlap in major features but only one of whom possesses the exact combination necessary to elicit a correct identification.

The implication of the Turing test results is that the real patient produced the kind of predicted utterance interaction patterns that were programmed into the computer and that this is why they were also diagnosed as paranoid. If all of the paranoid patients (and it is a little unclear from the report whether there were many or one) produced interactions congruent with the Colby theory, then we are still in the same methodological blind alley, since we do not know whether diagnoses of paranoia might have been given to patients who exhibited different patterns. If, on the other hand, analysis of the responses of the real patients uncovers some who followed different interaction "programs," then we should expect them either not to have been diagnosed as paranoid or – if they were so diagnosed – we must regard Colby's theory as having been falsified. Lacking these data, it is not possible to do more than suggest that such analyses might be performed and reported before we can assess the status of the theory.

### PARRY and the evaluation of cognitive models

James R. Miller

*Department of Psychology, University of Colorado, Boulder, Colo. 80309*

PARRY, in its various incarnations, has probably started more arguments among psychologists, AI researchers, and philosophers than any other single project in recent memory. A particular language comprehension system might be called implausible or ad hoc, but never "immoral" or a "patent fraud," as has PARRY. That PARRY has been argued about for almost twenty years now suggests that (a) there is almost certainly something intellectually important going on here, and (b) there is little to be gained by my entering the fray: no one is going to settle this issue here, and I am sure that both critics and supporters of PARRY will be well represented in these commentaries. Instead, I would like to focus on two other aspects of PARRY: its structural richness and its potential validity as a model of paranoia.

One feature of PARRY is particularly worth emphasizing. Even if PARRY should be shown not to be a valid model of paranoia, it would still be important as a case study of the number and variety of cognitive processes that are part of language comprehension and language use. One of the major points that has been raised in cognitive science recently has been the importance of intentionality and metacognitive structures to intelligent systems. PARRY certainly exemplifies

these issues, and has in fact been concerned with them since its inception.

My second point concerns how and whether PARRY can be verified as a valid model of paranoia. As Colby acknowledges, the ability to pass the Turing test is a necessary but not a sufficient part of the validation process. In my own area – prose comprehension – similar identifiability techniques have been used, primarily paraphrase and summary generation (Schank & Abelson 1977) and the production of reading time and recall data that are comparable to those of humans (Miller & Kintsch 1980). However, once these initial tests have been performed, it is necessary to dig deeper into the model, and extract increasingly specific predictions about increasingly small parts of the model: different parts of the model may be either good or bad, and these lower-level tests can help to identify which parts of the model need work. This has to some extent been done by Colby and his coworkers (Colby, Hilf, Weber & Kraemer 1972), but I suspect that these evaluations will need to be far more detailed before anything final can be said about the status of PARRY as a valid psychological theory.

Correspondingly, we might consider some of the ways that theories can be evaluated, and judge PARRY from those perspectives.

**Does the model predict data?** Yes and no. The quantitative validation of PARRY has mostly been limited to the Turing verifications described by Colby; more detailed analyses are needed. Just as the relevant verification of a psychological model of problem solving would go beyond simply determining whether the model was capable of solving the problem, and would include measures of whether the model used the same solution strategies at the same times as did human solvers, data suggesting that statements capable of triggering paranoid responses from PARRY have similar results on human paranoids would be useful here. Similarly, statements that improve PARRY's emotional state should affect patients in much the same way. This sort of verification may be difficult – psychiatric interviews are much harder to control than problem-solving experiments, and human-subject-use committees may hesitate to approve research programs that intentionally try to elicit paranoid states or are based on the recommendations of computer programs.

**Is the model generalizable and falsifiable, or is it simply an ad hoc description of its domain?** PARRY does not score particularly high on this criterion. In some respects, this criticism is irrelevant, since paranoia may not be a specific instance of a more general psychiatric problem. Naturally, if there were such a generalization-instance relationship here, the ability to extend PARRY to these other domains would be quite significant. PARRY unfortunately has a number of features that would at least impede this generalization. Most significant of these is the explicit way that the general belief system (for example, high shame plus interest by the doctor lead to a negative interpretation of this interest) is interwoven with the specific beliefs possessed by this instantiation of PARRY (for example, fear plus the belief that the doctor wants to do him harm lead to the belief that the doctor is a gangster). Certainly, some specific information must be included in the model in order to allow the conversational use of PARRY, but, from the viewpoint of extensibility, it would be better to have the identification of the doctor as a gangster grow out of a more general set of beliefs. I do not mean to say that it would be impossible to generalize PARRY in this way, but we should remember that this sort of generalization has been the downfall of more than one AI system.

**Does the model make sense out of a seemingly incoherent set of facts? (and) Does it make novel (and correct) hypotheses that countermodels do not?** These are important questions, and if they cannot both be answered now, at least some of the blame should be placed elsewhere. The first question is

easily handled by PARRY: the part of the model that leads to paranoid behavior is quite straightforward, and it is impressive that so simple a concept can, at least by Turing test standards, explain paranoia. The problem comes in attempting to compare PARRY to countermodels – so few of these exist at a similar level of explicitness that it is difficult to judge whether PARRY really predicts anything about the cause and treatment of paranoia that other reasonable models would not. Specifically, PARRY's recommendation to "minimize shame-induced distress" might be an over-specific part of a more general and correct treatment: it might generally be successful, but so might other similar (but related) treatments. Alternatively, other theories might take a different approach to the sources of paranoia, such that the events identified by PARRY would still be critical to the onset of paranoid behavior, but for reasons other than producing shame. Such differences could lead to different recommendations for treatment. Again, the need here is not so much to criticize PARRY, but to urge the development of equally specific countermodels, so that we can have some alternative explanations to consider. The flaws in one model are often not obvious until a competing model exists for comparison, and in a domain as complex as paranoia a strong explicit model is needed to suggest, constrain, and guide a coherent experimental program of research.

On the whole, however, I can find little to argue with here. Colby has set himself a demanding task, and has come away with some remarkable insights, in terms of both artificial intelligence and psychological theory development. I should also emphasize the "weak AI" nature of this work (see Searle 1980) – PARRY is an instantiation of a particular theory, rather than an attempt to create machine intelligence (or craziness). The fears cited by Weizenbaum (1976) and others that there is something wrong or even sinister going on here can be discounted. Psychiatric treatment based on experience with PARRY would simply reflect the emphasis on shame reduction postulated by the theory underlying PARRY. At this level, the computer on which the theory is instantiated is completely irrelevant.

## AI and cargo cult science

James Moor

Department of Philosophy, Dartmouth College, Hanover, N.H. 03755

In "Modeling a Paranoid Mind" Kenneth Colby raises a question designed to send chills down the spines of funding agencies. Is artificial intelligence (AI) cargo cult science? Of course, there is value in AI. If nothing else, playing with computers is fun. But can the time and money invested in AI be better spent? Will the cargo be rich enough? I think that it will be, but there are a number of methodological issues raised by Colby that, if taken seriously, will improve the chances for the success of AI.

Colby distinguishes between models and theories. I believe it is important to distinguish among programs, models, and theories (Moor 1978). A program is composed of computer instructions, and instructions are imperatives that are neither true or false. A computer model is made up of computer features such as programs, data structures, inputs, and outputs. These features are part of a model inasmuch as they are regarded as analogous to a subject matter being modeled. In an extreme case the computer model may be nothing more than the output, as with fancy computer graphics. But in AI the internal story is important and some of the program is usually relevant to the model. For the sake of the model, collections of programming instructions are taken not as sets of imperatives but as processes. Contrary to a popular opin-

ion, one does not give a model by simply listing the program. The AI researcher must explain which computer features are relevant and which are not. A given program can serve many models, and a given model can be constructed from many programs.

A theory is a system of statements that are true or false and describe the entities, processes, or events of a subject matter. A computer model is not true or false but has analogies with some of the entities, processes, or events described by a theory. Thus, a theory can have many computer models, and a computer model can be a model of many theories. Colby says, "it should be possible to state a theory independently of its model, so that those who do not understand the programs can understand them in a nonprogramming language." I say that it should be possible to state a theory independently of its model so those who *do* understand the programs can understand them in a nonprogramming language.

When programs, models, and theories are being developed, the connections among them are often loose. The theory may evolve out of the model that evolves out of the program. But at some point the connections should be made clear. Here Colby sets a good example by pointing out what is included in his model and what is not. Because the relation among programs, models, and theories is a relation of many to many to many, there is a great deal of room in AI for doing cargo cult science.

To what extent are computer models scientific? AI researchers often emphasize the explanatory power of the models. Colby states, "Such an explanation unifies observable input/output patterns by hypothesizing an internal structure of effective computational patterns connecting themselves and the input/output patterns into a comprehensive and organized whole." A model as a realization of a theory explains the computer's behavior, but why believe it has explanatory value beyond this? The usual response is that the model shows how things could work. However, this is not a sufficient answer. AI gives possible explanations, but science demands actual explanations. A model has little explanatory power until its associated theory has been confirmed independently. This is the dilemma for AI projects involved in cognitive simulations. If researchers are content with doing only model building, there will not be severe tests of the theories realized by the models. In this mode, AI takes on the look of cargo cult science. On the other hand, if researchers test the theories independently of their models, they are no longer doing AI.

The solution to the dilemma is to adopt a broad conception of AI that encourages testing beyond building models and running programs. AI research should be aimed at prediction as well as explanation, for correct predictions supply credibility to explanations. One obvious prediction from a simulation model is that the input/output behavior of the computer is similar to that of a human with regard to the dimensions in question. Turing-like tests, such as the ones Colby uses, provide a good way of testing such a prediction. As much as I would argue for the power of a Turing test (Moor 1976), I think there is another test of a model that is more impressive. The advantage of having the model on a computer is that one can manipulate the model. Such manipulations should suggest novel behavior patterns or lead to unexpected output that can be compared to human behavior. For example, suppose that the model PARRY suggests behavior that should be characteristic of paranoia but has not heretofore been noticed as such. If this predicted behavior is discovered to be an identifying mark of paranoia in humans, then there would be striking confirmation for the theory realized by the model.

For AI the difference between cargo cult science and science is the difference between writing programs to generate possible explanations and articulating theories to be tested as real explanations.

## Modeling paranoia: The cargo cult metaphor

Keith Oatley

*Laboratory of Experimental Psychology, University of Sussex, Brighton BN1 9QG, England.*

I was struck by the somewhat subdued tone and the unusual honesty of the reflections here that doing AI might be a kind of "cargo cult science." Colby has done many of the "right" things in his modeling of a paranoid conversation partner. His graduate students have earned their PhDs. His program passes Turing's test, which despite the "canned" speech productions is impressive. Yet still the question remains whether the aeroplanes bearing their cargo of scientific insight have been spotted.

I take it that a major goal of AI in psychology is to make explicit the mental processes of inference of which we are unconscious. I take it too that Colby's specific objective is to understand how remarks and events are interpreted by a paranoid person. The more general objective on which the specific one must be based, is to represent defensiveness so that we can understand the structure and implications of those defensive (that is, only partly successful) strategies which most of us have, and which include assumptions that other people might harm us in some way.

For both the specific and general cases, two senses of the term "unconscious" are relevant. One is Helmholtz's usage: only being aware of conclusions of mental processes, never of the means by which they are reached. The other is Freud's sense of the unconscious as a strategy by which we act while keeping from ourselves (and usually less successfully from others) some of our goals and plans.

It is at "unconscious inference" in Helmholtz's sense that AI work has primarily been aimed, and this is represented in Colby's program by the parsing module, and by most of the interpretation-action module. The simulation is important here in showing how conversational replies may be embedded within a structure of purposive behaviour in which utterances are interpreted as speech acts, and interaction is in terms of a script.

Part of Colby's originality, however, is to attempt to address the problem of "unconsciousness as strategy" as well. What we learn here about the unconscious strategy of paranoia comes from the goals and plans implicit in the interpretation-action module. The major goal is to avoid increments of the shame parameter attached to the self-definition beliefs. The plans associated with it, which are activated by high levels of the distress parameter (which is a function of the shame parameter), are (a) to attribute the cause of such increments to others, and (b) to respond to an interlocutor with hostility.

Only now, after all the computational stage setting of the "unconscious inference" kind, can questions like, What kind of strategy is paranoia? really begin to be explored. One can now imagine asking what (other than time delay) might cause the shame parameter to be decremented. Or is the central hypothesis about paranoia that shame cannot be reduced except by delay? One could ask, as Colby does, about how to represent conflicts - though it seems possible that before such a sweeping reorganization of the program, which Colby says would be necessary for this problem, one could ask what would be the effect of providing the model with plans other than hostile response - for instance, being conciliatory or evasive. Some paranoid people can mollify and evade, and such plans are just as defensive; but would the model still be seen as paranoid by psychiatrists if it used them? What would be the effect of having the model engage in active rather than reactive strategy, changing its plans in response to the amounts of shame elicited by the interviewer?

It may not be possible to answer such questions at all easily

using the current model as a test bed since, as Colby says, adding new rules is likely to be disruptive; but they do have to be answered if computational work is to show convincingly that paranoia is well defined and is a "skewed" version of ordinary interaction. Perhaps it will be some understanding of the principles of interpersonal strategy under assumptions of imminent attack that the hoped-for aeroplanes will bear.

Alternatively, perhaps the two kinds of understanding corresponding to the two kinds of unconsciousness need to be distinguished still more sharply. Unconscious inferences of perception and language evidently can be understood in terms of computational analogies. But if Helmholtz was right, we will not be able to affect them in any way by understanding them. On the other hand, it is perhaps the insight into cognitive processes that we might be able to change by understanding them that some workers in AI and psychology are seeking. Specifically in Colby's work this amounts to the hope that understanding might lead to the possibility of therapists or patients being able to see more clearly what is wrong with defensive strategies, thus allowing them to be debugged more effectively.

If this is the kind of insight that we are seeking, it seems possible that Colby's program provides a basis for the exploration of the relationship between goals and plans, and the factors that make them either rigidly unresponsive to events or capable of being debugged. This kind of exploration might indeed start to bring the odd insight-bearing plane down towards the airstrip.

## The dichotomous predicament of contemporary psychology

V. Pinkava

*Psychology Department, Severalls Hospital, Colchester CO4 5HE England*

Working AI models for all kinds of psychological phenomena are always a remarkable achievement, requiring much more ingenuity, special skill, and application than other kinds of theoretical work in the behavioural sciences. It is therefore with some regret that I feel I have to be slightly critical of the whole approach in this particular paper. What it seems to me to illustrate best is the fact that contemporary psychology in general, and abnormal psychology in particular, still seem to be dominated by the two inescapable spectres of the past: a general motivational approach on the one hand, and behaviorism on the other. We seem to carry on happily and implacably in these inveterate modes of thinking in spite of the fact that nowadays we do possess means and potential insights for moving to greener pastures along the road provided by the theory of self-regulating systems.

Colby's target article shows beautifully how much we all tend to be trapped inside traditional conceptions and approaches even if we try consciously to move ahead to something more recent and productive. His model of a paranoid mind is basically motivational: it infers that "paranoia" as Colby understands and defines it derives from a particular (maladaptive) strategy for dealing with distress induced by the affect of self-humiliation. But the author also states explicitly that "why only some people are susceptible to developing paranoia from shaming socialization remains a mystery."

Thus, the etiological theory behind the model is clearly motivational and vaguely "dynamic" (in the usual psychiatric sense): there is nothing basically wrong with the patient's brain or mind in terms of "hardware"; what is wrong is only the maladaptive programming, creating "paranoia." If the

patient could be "reprogrammed," that is, if we could influence him to drop this strategy, he would be cured. Therefore, the model is "mentalistic" in this sense, and also unrealistic in the experience of noncommercial psychiatry.

As for the concrete shape of the conception, as embodied in the program, one should note that the approach under scrutiny does not attempt to solve the problems of why or how a paranoid state has come about. The model uses the parser approach to create exemplary English-Language communication, while the "shame" parameter is just introduced ready made.

It seems that what this model of paranoia shows is basically the ability of a parser system to be made to act as a paranoid person - something akin to making an actor talk on a given theme, or in a given style. It seems plausible that, at least in principle, one could construct other working models of paranoid verbal behaviour using a different theoretical background, for example, the classical Freudian one concerning suppressed homosexuality. It might create greater technical difficulties, but such a model would be equivalent to the shame-based one both in performance and, I am afraid, also in heuristic merit.

The Colby model does not support or refute any theory of paranoia (as he defines it); it only shows that, with sufficient programming ingenuity, one can make a computer credibly behave as a paranoid patient. Colby may argue that it can be worthwhile to reproduce "the taste of the soup" in science instead of giving the recipe, but it is questionable whether the formulation of a set of assumptions in the 'active,' flexible form of a computer program is doing more than simply stating the principles along which one intends to set up such a program. The proof that the taste of the soup can indeed be replicated is required only by those who underestimate the skill of the hypnotist, or in this case, the programmer.

In the present commentator's opinion, it would be much more interesting if a program were written such that it generated paranoid behavior by an interpretable flaw in the program logic, rather than by the instigation (in program form) of a preconceived ready-made motivational principle formulated along very traditional phenomeno-mentalistic lines. In other words, the cause of the paranoia of the model is the programmer at his keyboard. The paranoia does not arise at run time from an initially 'healthy' data base by anything other than an ad hoc process (one designed to produce precisely this). The lifelike end is then taken as justifying the means, at least at a descriptive level.

As long as the program remains compelling to the diagnostician, it can be considered an equivalent, an analogy, or a very good simulation, but hardly a model in its strict sense, as in mathematical logic (Grzegorzczuk 1961). It is limited in that it can never show behavior that would not be expected of it by someone familiar with the program. In that sense it is artificial, and not intelligent at the root. The fact that the principle of paranoia operating in the model is ready made and put there in this form decreases its heuristic value. The fact that this is a motivational model seems to me to indicate that it is nothing very new. It follows the well-trod path of the post-Freudian (and actually pre-Freudian) phenomeno-mentalistic approach to psychiatry in a new guise.

Pre-Freudian psychiatry tried to explain mental disorders by a kind of "pseudo-causality" at a phenomeno-conscious level. A patient behaves this or that way because he has a wrong idea in his otherwise intact mind. If we make him abandon this "insane" idea he will become "sane." In principle we can do this in the same way we do with a sane person: by persuasion, coercion, and so on. It is only more difficult to achieve. Yet, however hard they tried, this never worked. Along came Freud and said that the wrong idea could not be removed just by reasoning with the patient, because the cause

of this idea was suppressed subconscious motivation which first had to be removed by bringing it to consciousness. This doesn't work either, despite repeated claims to the contrary, motivated, one suspects, by extrascientific reasons.

The idea then arose to look for some kind of 'mechanical' subphenomenal causes of psychological phenomena and their impairments, but to date this has not moved much beyond an oversimplified theory of learning. This is both regrettable and incredible, because nowadays we have the necessary theoretical armory to begin understanding psychological phenomena just as we understand the principles of the behavior of man-made information-processing devices. In the meantime, the "psychodynamic" psychopathology has grown less far-fetched and tends once more to approach the 'conscious level' motivational theories of the 18th and the 19th centuries. An interpretation of Colby's program as accounting for paranoia seems to belong there, in spite of its involving the latest in systems technology, as a tool! This is why I started this commentary by saying that we tend to be trapped inside the gravity field of motivational psychology. The reason Colby's model is considered to fit is its fidelity in input/output terms: taking this criterion for fitness as sufficient reveals that we are still trapped in the behaviorist universe. The Colby approach is a remarkable blend of irreconcilable ideas: a mentalistic conception is tested in a behaviorist way.

I do not want to expatiate on the question of input/output fidelity and the value of Turing's test: programs can be designed to demonstrate that humans are rather easily fooled into attributing some measure of intelligence to conversational programs operating at a relatively simple, "mechanical" level. This could perhaps be a description of PARRY, but it refers in particular to programs that select their replies at random from an inert data base. Implausible as it may seem, such a totally unintelligent program has had marked success in deceiving the human user (Pinkava 1980). The restricted universe of a man-console conversation is not test enough of a machine's credibility as a human parable.

One might also disagree on some points of definition and characterisation of paranoia as given by Colby, although from a more global point of view this may be less important. It seems that the classical distinction between paranoia as a nosological entity, and paranoia as a syndrome or subsyndrome has been lifted. This does not seem to be conducive to conceptual clarity.

The Colby model surely has its merits as a successful simulation in input/output terms, but in my view, its heuristic importance is rather restricted.

## Psychiatry and computers: An uneasy synthesis

William H. Reid and John F. Riedler

Nebraska Psychiatric Institute, University of Nebraska Medical Center, Omaha, Neb. 68106

In "Modeling a paranoid mind," Professor Colby uses syntactic structures and transformational-generative grammar to convey the concept of metalinguistics to be the reader. One observes the model by, in a sense, interfacing with it.

In spite of the author's choice of a symptom complex with high reliability of rater recognition, and his clear definition of paranoia as "a causal chain of strategies for dealing with distress induced by the affect of shame-humiliation," the model generates stereotypic behavior. Such behavior seems unlikely to be more than a visible-expressive part of the paranoid whole (a concept that need not detract from its usefulness in certain research settings, however).

Shame is in the eye of the beholder. The internal, symbolic, highly personalized view that each of us has of his self and environment goes beyond Colby's mention of "trivial . . . volumes of facts." The model is a mimic, apparently functionally similar to "paranoia" as tested, but (arguably) unlikely to be structurally comparable. Colby raises this issue briefly late in the paper, in part as a sort of "straw man." We wish he had focused on it more. The intellectual's fervent wish to codify affects and feelings cannot be made to come true by more and more obsessive reduction of some of their elements.

In saying that he and his colleagues are modeling the way a patient responds to input, Colby introduces a disclaimer as a premise. If one accepts the premise, then acceptance of the rest of the paper follows more easily. Eventually one is led to the seemingly rhetorical question, "Would it be a mistake with serious penalties for psychiatrists to think of the generation of paranoia in patients as similar to its generation by the relations and structures that characterize the model?"

We do not find the question quite so rhetorical. The answer is yes. Some clinical characteristics of paranoia have become the building blocks for the model. To compare patients to the model would be circular reasoning, at best adding an extra process to the clinician's efforts at understanding his patient.

Clinical observations → diagnostic formulation  
vs.

Clinical observations → computer model  
→ diagnostic formulation

This, in Colby's words, might lead to no "serious penalties" in patients being *studied*; however, the differences between diagnostic impressions that are limited by a function of model parameters, and those limited by mature clinical judgment of a unique patient would seem potentially critical for the proper diagnosis and eventual well-being of patients who are being *treated*.

Colby's issue is certainly valid, and his (and his colleagues') design admirable. The execution of the design is a mammoth task whose bulk, as implied in the last paragraph, becomes unwieldy even in the study model. We submit that in order to transcend sophomoric, global impressions of paranoia a *clinical* model would require such bulk as is, at present, the province of only the human mind and the combinations and permutations of its developmental experience.

## Evaluation of a model's test

Russell Revlin

Department of Psychology, University of California, Santa Barbara, Calif. 93106

Colby's enterprise amply illustrates the difficulties in evaluating any model - simulation or otherwise - with respect to criterial behavior that is not quantifiable. Nearly a decade of research in this area provides at least one unequivocal result: the Turing test cannot be taken seriously. Its use is invariably subject to methodological and inferential criticisms (see below), and its purpose is motivated by historical issues that should no longer be taken seriously. Psychology's concern should not be to joust with those who challenge the computer's ability to simulate thought; rather, we should get on with developing those models that accomplish the simulation, and we should evaluate their ability to capture relevant generalities and account for specific behaviors. While the computer is a marvelous vehicle for such an enterprise, "hand simulations" may be no less strenuous or stringent and may even be more efficient (for example, by obviating the need for complex parsing systems). Whatever the vehicle, we must

keep in mind that the model is our primary focus and it should be assessed against criterial measures, be they reaction times, decisions, or counts of accusatory statements. In the absence of strict criterial measures and point predictions, we need guidelines for the "reasonableness" and "suggestiveness" of simulations to avoid the difficulties of null-hypothesis testing that are characteristic of Turing-like procedures.

At least four criteria for reasonableness of models and their evaluation are prompted by Colby's tenacious efforts (see, for example, Colby 1963; 1973; Colby, Hilf, Weber & Kraemer 1972). These are (with apologies to Grice 1975): informativeness, truthfulness, relevance, and transparency. With respect to *informativeness*, models ought to be comprehensive in accounting for the criterial behaviors as well as illuminating underlying mechanisms that may be generalized to new situations and theories. For example, Colby's model is intended to simulate the responses of a paranoid patient during an interview. Such a model would be informative if it could make predictions concerning the effects on the putative patient of different forms of therapeutic intervention. The model might then be less ambiguously evaluated by determining whether the predicted effects occurred in the simulation as they presumably would with a human patient. In sum, we are informed when the model makes a prediction whose evaluation is open to inspection and may even be quantifiable.

Evaluation of models ought to be *truthful*. If the intention of a test is to see whether an interrogator can distinguish between a computer and a noncomputer patient, then the judges ought to be trying to make that distinction. Colby et al. (1972) do not inform their judges beforehand that a simulation is involved. The "tests" then do not measure the judges' ability to distinguish between the "patients," but rather test whether interrogators can make allowances for departures from normal, human interactions because they believe that they are examining a human (see McCorduck 1979). While such a procedure is in keeping with the Turing test, it would appear to bias the evaluation in favor of affirming the null hypothesis - perhaps the most critical limitation of the test.

*Relevance* and *truthfulness* are related; the former refers to the presence of specifiable criterial behaviors on the basis of which the model will be evaluated and to the fact that the simulation of such behaviors requires that the model actually work. For example, if we get a cooperative answer (in Grice's sense) to a question, we assume that we have been understood and that the participants in a dialogue share a common objective. Paradoxically, if an answer in Colby's system is only peripherally related to the question, the interrogator assumes not only that the question has been understood, but, more important, that the reply comes from a paranoid patient. That is, the presumption of modest forms of noncommunication is that the model is valid (see evaluations of nonsequiturs in Colby et al. 1972). This renders the evaluation of the model problematic. However, if the criterial behaviors are specified beforehand (e.g., characteristics of the client-therapist interaction in addition to those accompanying normal conversation), then we have a kind of taxonomic checklist against which to evaluate the relevance of the "judges" decisions.

Finally, the models - if not their tests - ought to be *transparent*. Simulations using computers add a measure of opacity to any model, by virtue of required instantiating mechanisms and rudimentary coding. The fundamental processes ought to be accessible so that the evaluations of the model can be subjected to hand simulations and decomposition by critical evaluators. Colby has done nicely in this regard in a number of presentations of his theory (for example, Colby 1973). Transparency permits the coding of new client-therapist interactions, which could only contribute to the generality of the model (hence informativeness).

These comments should not be taken as indicating limitations of Colby's model. Beyond its value for psychiatric theory, Colby's work has been important in bringing into focus issues of computer simulation that were only marginally plausible conjectures when he and his associates initiated their enterprise. Nor does the foregoing argue against the appropriateness of the ratings of expert judges. Rather, I suggest that a particular form of such tests (attributed to Turing 1950 and Abelson 1968) has become reified beyond value, serving to obscure perhaps more sensible if not more illuminating evaluation procedures.

## Deep and shallow simulations

Aaron Sloman

*Cognitive Studies Programme, School of Social Sciences, University of Sussex, Brighton BN1 9QN, England.*

A deep simulation attempts to model mental processes, whereas a shallow simulation attempts only to replicate behaviour. The question raised by Colby's paper is, What can we learn from a shallow simulation?

The only clear answer I can find in this case is that insofar as the behaviour is verbal, the design of a relatively *robust* shallow simulation can contribute to our understanding of some of the difficulties of interpreting unconstrained sentences. In particular, it seems likely that this sort of project could unearth previously unnoticed features of English grammar, and pinpoint some of the problems of coping with incomplete, ungrammatical, or misspelt input. How far our understanding of these issues has been advanced is not clear from Colby's paper. In particular, it is not clear whether anything new has been discovered about how to integrate the processing at different levels (syntax, semantics, pragmatics, inferences, and the like) in order to achieve robustness.

As far as understanding paranoia is concerned, I suppose the program can be said to embody some sort of transition net, showing which changes of state can be triggered by various types of input. However, there is no real analysis of the nature of those states, or the processes of change. That would require a far more complex model, showing how beliefs and motives of various sorts are represented, and how they can interact with each other and new information to produce new beliefs and motives and various kinds of disturbances of processing which we recognize as emotions. For instance, feeling afraid involves a complex cognitive state involving beliefs about what is likely to happen and desires that it should not happen, together with disturbances of other cognitive and physiological processes. Being angry involves an even more complex collection of beliefs and interacting motives. Only when we have a deep understanding of such computational states and processes as they occur in "normal" people can we hope to understand what is going on in pathological cases. I shall shortly submit to this journal a long paper surveying some of the processes involving motives, deliberation, decision, and emotions in intelligent animals and artefacts. Perhaps such work can eventually be integrated with Colby's.

## Is PARRY paranoid?

David W. Swanson

*Section of Psychiatry, Mayo Clinic and Mayo Foundation, Rochester, Minn. 55905*

This target article reports an intricate and time-consuming effort at computer simulation of the paranoid mode of

thinking. PARRY is endowed with a truly extensive memory of words, idioms, and concept patterns, in which specific emphasis is placed on personal history, interpersonal relations, and a delusional system. Colby has succeeded in endowing a machine with enough paranoid wherewithal to confound unsuspecting psychiatrist interviewers. Psychiatrists are not too hard to fool, but nevertheless one has to be impressed with the resources and flexibility of PARRY.

Appearing paranoid, however, is not the same as being paranoid. With appropriate training a simulated patient can surely pass for someone genuinely paranoid. Some persons have, through deception, achieved hospital admission followed by the desired diagnosis and treatment (Rosenhan 1973). These intact, presumably nonparanoid persons may seem to be at least as paranoid as PARRY, but little is to be learned from them about clinical paranoia beyond their contribution to the development of skills of clinical acumen.

From a clinical standpoint, PARRY is a sophisticated paranoid. Included in his circuitry are components of affect and thought disorder, and they can interact to influence PARRY's responses. Increasing PARRY's versatility is the incorporation of an inference system in which PARRY develops beliefs about the interviewer and, by increments and decrements, matches them with those in his stored belief system. This is surely reminiscent of paranoid patients' efforts to explain their distressing feelings or perceptions on the basis of previous experience - which we all do - but then when unsuccessful, paranoid patients immediately project their causes outside themselves. Colby obviously understands paranoia and has programmed PARRY well. Similar efforts at educating a simulated patient would have produced an equally sophisticated model, but in either instance, it is still only a model.

What we seem to have is Colby, a dedicated student of the paranoid mode of thinking, putting his knowledge into the programs and hardware of PARRY; but I am not satisfied that moving it from the author's intellectually sophisticated, nonparanoid mind to PARRY's computer has produced a paranoid entity of any depth beyond that of an actor or a malingerer.

Perhaps what reduces my enthusiasm for PARRY's value is the belief that the paranoid mode of thinking is present in all humans and is used by most of us under certain conditions. In the clinical setting the problem is that it is used incessantly. It is unlikely that the crux of a person's paranoid thinking relates to the engrams of shame, guilt, latent homosexuality, or what have you, and how they are provoked by that person's interaction with others. The paranoid mode of thinking is an epiphenomenon. The more fundamental issue is why a seemingly intact person will oftentimes rather abruptly become paranoid and persist in thinking that way, resisting reassurance and realistic explanation of his paranoid way. I am uncertain from the data presented whether PARRY can contribute to solving this clinical dilemma.

Does the combination of words, idioms, concept patterns, affects, thought disorder, and inference system built into PARRY produce something beyond the sum of the parts? Is there a higher level of integrative functioning that in some rudimentary fashion mimics his human counterpart? Colby is suggesting that this is true, and he will presumably be testing this hypothesis and providing us with additional data. I see more potential in the continued use of a model amphetamine psychosis (Angrist & Gershon 1970; Griffith, Cavanaugh, Held & Oates 1972) or a potential phenylethylamine model of paranoid disorder (Potkin, Karoum, Chuang, Cannon-Spoor, Phillips & Wyatt 1979); but nevertheless the possibilities of learning from PARRY are worthy of pursuit.

A goal in developing the paranoid model is its potential use in guiding pragmatic action. If minimizing shame-induced distress is effective in altering the paranoid mode in the

model, it might be useful in the treatment of paranoid patients. Although not everything can be included in a paranoid model, the computer simulation does convey a simplistic view of most paranoid disorders. The view is that if the patients can somehow recognize the role of shame, or factors inducing shame, they will overcome their disordered thinking. As Colby knows, this is likely to apply to only a few paranoid patients. What is more, this basic concept of PARRY's potential usefulness implies that should the interviewer avoid shame-inducing behavior, then normal thinking will gradually be restored to PARRY's human counterpart. It is impossible to tread this carefully with most significantly paranoid persons; and even if success is achieved by the interviewer, it may only reduce the paranoid mode of thinking in that specific dyad. That is, of course, worth something, and if PARRY can assist in training for this difficult task, he will be of some value. At present, medication is usually required to reduce the provoking affect in the patient's associated hypersensitivity.

#### NOTE

Copyright © 1981 Mayo Foundation.

### Clinical artificial intelligence

Virginia Teller and Hartvig Dahl

*Downstate Medical Center, State University of New York, Brooklyn, N. Y. 11203*

**Experimenter:** Thank you for agreeing to participate in our study, Dr. Teller. Your task will be to use this computer terminal to hold a conversation with a psychiatrist whose replies will also appear on the terminal. At the end of your dialogue you must decide whether you have been conversing with the real Dr. Hartvig Dahl or a computer model of him that we are testing.

**Teller:** OK, fine. [Sits at terminal and begins typing.] Good morning, Dr. Dahl. How are you today?

**Dahl:** Not great. I've got a bad head cold. How about yourself? Are you another one of these specialists they've called in to see whether you can figure out if I'm the real me?

[Exp: Note the old psychiatrist's trick of answering a question with another question.]

**T:** I guess so. If you're a computer model of a psychiatrist rather than a real person, you must have heard about Colby's PARRY. Tell me something about paranoia.

**D:** Well, for one thing, paranoia is one of the most easily and accurately diagnosed of all psychiatric ailments. All it takes, according to our latest set of diagnostic criteria, DSM-III, is "a chronic and stable persecutory delusional system of at least six months' duration." That's why PARRY is such a giveaway. I've had several conversations with him, and as soon as he mentions that the Mafia is after him, any first-year resident would peg him as paranoid.

**T:** Can it be treated?

**D:** Not that I know of. Freud said in 1922 that "cases of paranoia are, for well-known reasons, not usually amenable to analytic investigation" (p. 225).

**T:** Aha! So you're a psychoanalyst.

**D:** I didn't say that.

**T:** But you quoted Freud.

**D:** You think only analysts quote Freud? That's naive.

**T:** You seem a bit insulting for a real psychiatrist. Let's get back to the treatment question. Can paranoia be treated?

**D:** I already answered that: not that I know of. Unless you're also crazy. The craziness can be treated with medication, but not the paranoia.

**T:** How ironic! The PARRY program itself exhibits some of

the characteristics of the paranoid personality. It has difficulty recovering from error, it can't learn, and it's very hard to change without driving its programmers crazy. Are you that rigid?

D: Now who's insulting who? What's your racket, anyway?

T: I'm a generative grammarian cum computer scientist.

D: You're one of those Chomskyans?

T: I guess you could say that.

D: Then you don't believe in the Freudian unconscious. Even PARRY is motivated by unconscious emotions. But never mind, I suppose you think that I can't generate grammatical sentences on my own - that I'm just a bunch of canned phrases and stored quotations. Just like PARRY à Deux with his 1,800 bottled - I mean, canned - remarks.

[E: Shame 3.1, anger 2, fear 1.1, distress .5, interest 6, joy 8.]

T: You certainly are sassy. How did you and PARRY get along?

D: Not very well. PARRY interacts successfully only with a psychiatrist conducting a straight diagnostic interview, which I'm not very good at, as you can see. When PARRY's assumptions about interviews are violated, communication breaks down. I'm reminded of Cerf's [1973] report of the conversation PARRY had a few years ago with Weizenbaum's ELIZA'S DOCTOR, which plays the role of a Rogerian therapist. Part of it went something like this:

PARRY: I don't understand your motives.

ELIZA'S DOCTOR: Why don't you understand my motives?

PARRY: Shouldn't I be?

ELIZA'S DOCTOR: Why do you say that just now?

PARRY: That's just the way it is.

ELIZA'S DOCTOR: Suppose that's just the way it were not?

PARRY: So?

ELIZA'S DOCTOR: I am not sure I understand you fully.

PARRY: You harp on the same questions too much.

ELIZA'S DOCTOR: What does that similarity suggest to you?

PARRY: I don't get you.

ELIZA'S DOCTOR: Do you wish you did get I?

PARRY: It's about time you showed some interest in my feelings.

ELIZA'S DOCTOR: Oh? I showed some interest in your feelings?

T: Enough. This is beginning to sound like a six-part rricercar. Let's get back to basics. What, in your opinion, is the scientific value of Colby's seventeen-year plunge into the paranoid mind?

D: Maybe he hit bottom.

T: There you go again, being insulting.

D: Well, you can interpret it that way. What I mean is that it seems like a dead end to me. Didn't Colby say that a Turing test is a weak test of a theory? I think that's true, and I think it's a weak test of the shame-humiliation theory of paranoia. I've been taught to think more of jealousy and even unconscious homosexual wishes as the sources of paranoia. But then, of course, Colby is not *my* programmer.

T: Perhaps we need more PARRYs, each one built on different theories of paranoia.

D: That's it! Let's have less "argumentation" and more experimentation.

T: I couldn't agree more. however, I must confess that I find Clippinger's [1977] ERMA program, which simulates the discourse of a woman in psychoanalysis, more intriguing than PARRY. ERMA has class. Her computer psyche (oh, how we anthropomorphize our computing machines) embodies a multiple-agency theory of cognition. Separate components in the system act and react, interrupting each other to product "conflicts" that are expressed as disfluencies (stammers, hesitations, and the like) in output. Even though ERMA is too

withdrawn to utter more than one halting paragraph of discourse -

E: [Interrupts] Your time's up, Dr. Teller. Would you please tell us whether you think you've been talking to the human or the computer version of Dr. Dahl and give your reasons.

T: [Still typing] Well, with a sigh of relief, I'm happy to report I'm convinced this is the real Dr. Dahl, not the computer model. He's much too witty, clever, sophisticated, and knowledgeable, and his language is much too rich to be anything but human. I'm inclined to take Hofstadter's [1979] position: "Truly" intelligent behavior turns out to be whatever hasn't been programmed yet.

[Teller and experimenter leave room.]

D: Little do they know that I'm not *really* under their control; the real Hartvig Dahl is, as he was described in a *New Yorker* profile [Malcolm 1980], an "older, wiser, more benign and authoritative version" of myself.

## Author's Response

### PARRYing

Kenneth Mark Colby

Neuropsychiatric Institute, School of Medicine, University of California, Los Angeles, Calif. 90024

Apart from the usual riot of confusion between theoretical explanatory models based on analogy and replicas or clones, and apart from the inevitable (but salutary) noise of grinding axes, it is difficult to sort the commentators' remarks into themes. Hence I will proceed alphabetically, by commentator, interweaving relevant views of other commentators whenever possible.

Abelson goes after PARRY and notes that the model has not been "fully accepted in psychological or artificial intelligence circles." Allowing for the historically tedious resistance to emerging novelty, and for the NIH (not invented here) syndromatic jealousies of a fledgling field, it remains true that PARRY evokes mutterings of dissatisfaction in some quarters as well as circles. One reason some members of the AI brotherhood do not like the model is that it does not instantiate the skills of "intelligence," the kind of behavior the field is supposed to be concerned with. Instead, PARRY instantiates a type of psychopathology, a terrain whose problems, in turn, are unfamiliar to most scientific psychologists. In borrowing concepts and methods from psychology and AI, the model falls into an inter-field crevice all its own.

Abelson raises two questions: (1) Does the model really explain paranoia? and (2) Do its mechanisms generalize to anything else? Explanations involve a questioner asking an explainer a "why" or "how come" question. If the question is answered in a way that is satisfactory to the questioner, the understanding becomes relativized between the explainer and the questioner, who share some background knowledge or consonant coordinate systems. Dozens of criteria have been offered for satisfactory explanations. Among

them are: (a) Is the explanation relevant? The explainer must grasp the central concerns of the question. (b) Is the explanation correct? The explanation should have evidential support in the form of empirical observations, show coherence with other knowledge, and have sound lines of reasoning. (c) Is the explanation testable? As yet, there may be no supporting empirical appraisals, but *in principle* the explanation should be directly or indirectly testable. (d) Is the explanation complete? No explanation ever completely explains anything. Completeness here means that the questioner is satisfied at some level, for the time being. (e) Does the explanation have explanatory power? That is, it must not only be accurate, it must also unify a range of diverse phenomena. (f) Does the explanation have predictive power? Prediction means to announce a condition or event without foreknowledge. Hence an explanation may retrodict old facts, not used in the construction of the explanation, as well as predict new conditions or new items that turn out to be as expected. (g) Is the explanation simple? Simplicity is a matter of intelligibility based on what is familiar to the questioner. Complexity is a methodological rather than an ontological category. (h) Is the explanation deep? Depth means that the mechanisms posited are far from observation and hence considered more fundamental. (i) Is the explanation fruitful? That is, is it extensible to a new domains, leading to fresh starts, new ideas, new research?

And so on. One can add favorite regulatory standards for explanations, such as prior likelihood, originality, metaphysical consistency, and elegance. Each explainer and questioner has his own list, and the items are not all compatible. Thus, each man must decide for himself whether our theory and model "satisfice" as an explanation of certain aspects of the behavior classified as paranoid.

Abelson calls the AI world's attention to the importance of affect and intention in the model. If the cognitive sciences limit themselves to cognition, they will not develop satisfactory explanations of mental processes that obviously contain far more components than cognition. The carry-over of PARRY in psychology and AI would lie mainly in a further understanding of affect and intention. An effect in psychopathology might be a variety of systematic experimental investigations of paranoid behavior. Do paranoids *really* behave like PARRY in jumping on one phrase in a sentence, ignoring the fact that the rest of the sentence denies what they are sensitive to? Some do behave this way, because we have observed them doing so, both in standard clinical interviews and in teletyped ones. But more systematic evidence is needed. Now if we could just get a grant and get through the Human Subject Protection Committee . . .

Agassi is concerned about the problem of diagnostic agreement among psychiatrists. It is true that many of the diagnostic categories show poor reliability. But paranoia or "paranoidness" is not one of them, as stated in the target article in section 2.1. Depending upon what measure is used for "extremeness," I think psychiatrists consider paranoia to be a severe condition in terms of both its interference with normal function-

ing and its great recalcitrance to treatment. The terms Agassi uses - "paranoia vera," "*idée fixe*," "catatonia," "la belle indifférence" - have all but disappeared from American psychiatry. Forty years ago catatonic schizophrenia was the most common subtype of schizophrenia. Nowadays, if a case appears in the hospital, we bring in the medical students to observe such a museum curiosity. What has happened? (Hint: the same condition may now be called something else).

In the late 19th century, Janet, a French psychiatrist, used the phrase *idée fixe* to describe the "automatisms" of neurotics (not paranoids). A fixed idea may involve anything from the belief of a businessman that he must wash his hands hundreds of times a day to the belief of a woman that she will soon change into a man. A persecutory delusion can be a fixed idea, but the converse is not true. Agassi asserts that we have an idiosyncratic definition of paranoid conditions. Our definition can be found in the 1980 diagnostic manuals of both the American Psychiatric Association (DSM-III 1980) and the World Health Organization (ICD-9) and in Kendler (1980). (See replies to Magaro & Shulman.)

Paranoids can have justifiable beliefs in persecution. As pointed out in the target article, paranoid behavior tends to provoke hostile attacks by others subjected to it. Hence, in a given clinical case, the feelings of persecution may be partly justified and partly not. Sorting these out is no easy task. I'm not sure what Agassi means by "pass a normal person for a persecutor." We had judges interview many nonpatients as well as nonparanoid patients by means of teletype, and none of these individuals was identified as a paranoid.

It is good to hear from Carbonell that current parsers work like ours in attempting to formulate meaning from ungrammatical input. The early parsing systems were overly perfectionistic and unrealistic. In real life dialogues, people don't speak so good.

Early AI approaches to natural language by those brought up in the Mandarin tongue of the first-order predicate calculus followed the Euclidean tradition in hoping to start from a few self-evident axioms (rules) and then to derive thousands of further consequences (sentences). Extensive experience with a large data set acquired from hundreds of people interviewing the model taught us that it is more fruitful to start with the wide richness and variety of the data and then try to reduce them to a narrower domain manageable by patterns as well as rules.

Carbonell mentions programs that model ideologies. How to test these programs, using criteria other than face validity, remains a nettlesome problem. He is concerned with proving-as-correct distinct computational mechanisms for belief-guided behavior. The first question is whether they are distinct and in which respects the distinctions make a difference. If different, they may all be correct, since people have many alternative mechanisms for reaching similar ends. One cannot decisively prove or disprove any theory in science. The main questions are: What is the gain? Where do we go from here?

Gunderson provided us all with some good laughs years ago with his rock-box simulation, presumably operated by a wizard behind the curtain. (When

accused of being a bad man, the Wizard of Oz said he was a good man – just a bad wizard.) Like computers, rocks by themselves don't do much and must be *made* to do something. In his comments here, Gunderson makes clear that by "program-resistant" features of mental activity he does not mean features that are explanation resistant, or even explanation resistant in the form of a model. His concern seems to be with the conscious subjective-feeling aspect of mental activity, not with the underlying processes that produce these feelings. Now we bang into the interminable confusion between model-as-set-of-assumptions, such as a billiard-ball model of a gas, and model-as-replica, such as a three-dimensional size-scale model or a frog clone. When we say PARRY reproduces the "taste of the soup" we mean only in its I/O (input/output) behavior. (Notice that taste does not include the sight or feel of the soup.) The I/O pairs are observable phenomena that can be compared to the I/O phenomena of real paranoid patients in initial diagnostic interviews. The I/O pairs constitute the imitation; the mechanisms that produce them are *not* imitations with resemblance, they are *representations*. Because both an observable resemblance and a theoretical representation are simultaneously present components of models such as PARRY, an exasperating confusion of the two can easily occur.

Further terminological tangles result when Gunderson tries to differentiate artificial intelligence from computer simulation. He takes AI to be exclusively "straight" AI, as in making computers clever at chess and backgammon. But from the start there has always been a branch of AI interested in theoretical psychology, that is, in the way human minds work. Maybe those of us in this branch should stop calling it AI and call it "algorithmic psychology," referring to representing mental processes in the form of algorithms. I doubt that my suggestion will catch on, because AI is now an entrenched term, thrilling to some and menacing to others. Since Turing-like procedures test the goodness of the imitation, not the representation, we should give up the designation "Turing-like" also. No one seems to know what Turing's test really is or what it implies.

Suppose we want to "model" the pain in the toe that results from its being pressed with a noisy, cracking sound by a large, sharp-edged, brown, meteorite-laden rock from Gunderson's fiendish rock box. What does "model" mean here? Do we want to reproduce the subjective or explain it? Do we explain the subjective by transcending it and analyzing it in nonexperiential terms? Do we want the entity not only to say "ouch! my toe hurts" but also to show other signs of pain? What signs? Facial expressions? Jumping around? Holding its foot? Many pain conditions can be judged only through interview data. I hope neither Descartes nor Gunderson has ever suffered the agony of migraine, because the only way to make a diagnosis of migraine is through the *oratio obliqua* report of a certain type of headache. The verbal interview data are "read" by experts as signs having a certain diagnostic and prognostic significance. There is one way, however, to be sure the "model" not only "has" the pain but subjectively "feels" it as a raw feel, and that is to construct a human clone. But the clone is no longer a

simplified, explanatory model. It is a replica. We should not carry lucidity too far, but in the interests of conceptual hygiene, we should keep clones and explanatory models separate.

What is illuminating to an explainer may not be illuminating to a questioner. The role of affects as "interrupts" in mental processes is illuminating to both Simon (1969) and ourselves, but not to Gunderson, who dismisses it as unilluminating. Theories can explain singular events as well as laws. When I explain to a surprised and perplexed psychiatry student that his paranoid patient interpreted a standard innocuous interview question as an attempt to humiliate him, which resulted in a sudden rise in shame, which in turn interrupted the patient's normal interview responses with the surprising production of a seemingly-out-of-context insult directed at the student personally, the perplexed student finds this explanation of an otherwise-puzzling event quite illuminating, in the sense that the explanation of the interrupt makes sense. But Gunderson presumably would not. Then what explanation of this chain of interview events *would* satisfy him? What more does he reasonably want at this time? We must await the revision of *Mentality and Machines* (Gunderson 1971) for illumination.

Izard & Masterson ask for references to the controversies about Turing's test. There are now thousands of papers on this test in the literature of philosophy of science, artificial intelligence, and the cognitive sciences. The references Colby 1975; Colby, Hilf, Weber & Kraemer 1972; Gunderson 1971; Heiser, Colby, Faught & Parkison 1980; Hofstadter 1979; 1981; and Moor 1976 would be good places to start. My own view is that we should give up further debate about Turing's test. Better interview tests should be designed, keeping in mind that when a model passes one, diehards can always claim it is a weak test.

Like Abelson, Izard & Masterson stress the important role of affect in psychological models. As far back as the turn of the century, there were case reports in the German clinical literature of paranoid conditions following closely upon shame-humiliation experiences. (For a review of this literature, see Retterstol 1961.) To note that a class or kind of antecedent correlates with a class or kind of psychopathology is the first step. The next step is to try to spell out a theory of the intervening internal processes that makes sense out of the correlation. It is to Tomkin's (1963) credit that he began this type of spelling out of paranoia (in literary rather than algorithmic form), and we have said so in many of our publications. The role of affects in our theory derived from Tomkin's work. It is true that we have introduced modifications, an inescapable consequence of a good working theory.

The field of affect theory still struggles to erect a satisfactory taxonomy of the concepts to be employed in the theoretical structure. Tomkins divided nine primary affects into positive and negative groups. In contrast, we posited enjoyment ("good," "pleasant") and distress ("bad," "unpleasant") as the two basic and primary affects observable in the first days and weeks of life. We posited that other, more specific affects become differentiated later in development and are capable of activating these two primary affects. (I

should mention that we have had many pleasant and invigorating personal discussions with Tomkins about these matters along the way.)

We well realize that shame is connected to contempt from others, that one can be ashamed of being ashamed, and that a parent can use contempt to evoke shame to control fear in a child. In real life, the interactions and their feedback amplifications and attenuations can all be dizzyingly complex. In the model, we wanted to get as much as possible out of as little as possible by having a limited set of cause-effect relations. Also, we wanted to make a clear distinction between paranoia and paranoid schizophrenia, as is increasingly justified by the evidence (Kendler 1980). When Tomkins wrote of shame-humiliation and terror in patients with delusions, he was thinking of paranoid schizophrenia (Tomkins 1963, p. 480). Our patient category is that of paranoia and not a subtype of schizophrenia. (The point is discussed further in our reply to Magaro & Shulman.)

In our theory and model, distress is *not* simply equivalent to intense shame-humiliation. Distress is the basic undifferentiated negative affect. Several other affects or affect combinations can influence it. Hence, in the model, anger, shame, and fear can combine to produce extreme distress. This is what is meant by emotion-induced emotion rather than emotion induced by the activation of a belief with specific semantic content.

A theoretical model can serve as an instrument for its own improvement. In thinking about the model and its theory, we were not quite satisfied with our own classification of shame as a differentiated affect on the same level as anger and fear, as mentioned in section 4 of the target article. Both in developing later in life, and in being more complex in its interrelations, shame seems to us to belong to a higher-level value system more closely tied to one's worth as a person than to biological survival. Babies don't blush, and young children do not develop paranoia. In time, this may all get unraveled so that we will eventually get it right.

The generality problem arises again in Kochen's commentary, but not in the sense of generalizing the mechanisms to other domains. Kochen points out that the model "cannot behave appropriately outside of its restricted domain of discourse." This is, of course, quite true and is one of the characteristics that make *any* theoretical model a model and not a complete facsimile. Simplifications lead to circumscriptions. PARRY is not a "patient" except in loose laboratory jargon; it is a model, a description, a set of strings of symbols. The front-end parser can deal only with a limited subset of English, and the internal patterns and processes can deal with only a limited number of psychiatric interview situations.

PARRY has built-in generalizations about doctors and what to expect from them in an interview. As discussed at the end of the target article, it would be interesting to have the model generalize about the hundreds of interviewers it meets over time and to modify its beliefs. Its "learning" and change would then extend over interviews rather than be confined to a single interview. The mechanisms for doing this are theoretically known and they have been implemented,

by Kochen among others, "to some extent." This extent is so small, however, that there must still be some difficult implementation problems involved or we would have a "generalizer" by now in more than toy-size form. The model's lack of a learning ability bears on our own main criticisms of the model, namely, that it is still at too high a level, for example, the representations are as belief propositions rather than single concepts. It should be able to construct molecular beliefs from concepts at a more atomic level.

We appreciate Kochen's appreciation of the parser's information-engineering skills. At least one of the "nonsequiturs" he reports deserves an explanation. It illustrates how a program's refined cleverness in desperate situations can lead to trouble. Kochen reports that the model was interviewed by a Dr. Chen from China who was born in Shanghai and educated in Peiping. PARRY believes what it is told, at least at the start, but it knows nothing about the proper name "Chen," nor about China, Shanghai, or Peiping. What to do? When it does not recognize words, one strategy the parser uses is to respell the words since, from hair-tearing experience, we have learned that most people do not spell very well over the teletype. We reran Dr. Chen's interview with a full trace and found that the parser started juggling "Peiping" by chopping off the "ing," rearranged the rest of the letters and came up with the dictionary-recognized word "pipe." This particular interviewer, Dr. Chen, instead of saying one thing at a time, typed in two sentences and a question. The model understood the "born" sentence (except for the geography) but ignored it as irrelevant to its concerns. In the second sentence, it comes up with "pipe," thinks something about smoking is being referred to, and responds blindly. The result is a horrible clank that should be a lesson to anyone who plans to allow unrestricted input in an interview.

Lindsay grasps well the necessity of an interface that must deal with these formidable parsing problems. Since clanks make the model look bad, why have a parser at all? But then how can we have the model plausibly judged to be a member of the natural kind at issue when, in clinical reality, membership in that kind must be judged on the basis of verbal behavior in any interview? The "talking" part of the model clouds all sorts of issues, such as the above-mentioned confusion between I/O pair imitation and postulated internal processing, which we claim have structural equivalence.

The ancient theme of the whole rising above its parts appears in Lindsay's mention of objecters to programmed psychological models: "they will yet argue that paranoia is of necessity a property of a total personality, and cannot be abstracted from such an embodiment and studied separately." This objection is not Lindsay's own view, but he wonders how it is to be answered. The holistic cry of what cannot be done has been yelped forever, but somehow it has not held back the accumulation of knowledge about what we want to know about. All my professional life in medicine, psychiatry, and psychology I have heard this tiresome admonition to wait (usually accompanied by a finger wag). "You cannot understand the brain until you understand the whole body; you cannot understand the

language area of the left hemisphere until you understand the whole brain; you cannot understand paranoia, or depression, or transsexualism until you understand the whole person." (There is more on holism in the reply to Reid & Riedler.) These are simply romantic laments or ideological proscriptions from naysaying people who like to issue orders. None of them is an established law such as those we have about the impossibility of perpetual motion machines. In the target article, I specifically point out that in our theory the paranoid mode is a bulk, or net-behavioral, property of the whole (postulated) system of the model's components and not just an isolated part.

Lindsay's regret at our lack of a belief-settling success rule in behavioral sciences is echoed in many other sciences. By nature, the scientific enterprise is forever unfinished. Poincaré noted even in mathematics there are no solved problems, only more-or-less solved problems. When you start to solve a problem, you soon realize you are unfolding a series of problems having no terminus. A success rule in medicine is usually a pragmatic one. That is, a man with a successful theory can *do* something. He can make something (often new) happen, or prevent something from happening, whereas a man with a rival theory cannot. When multiple candidate theories exist, they may all be correct, representing alternative causal pathways toward the same end. What produces malaria? (a) bad air, (b) swamps, (c) stagnant water, (d) mosquitoes, (e) certain kinds of mosquitoes, (f) females of certain kinds of mosquitoes, (g) parasites, (h) parasites that can be killed by chemicals. Success would have accrued to each of these theories if their holders could have manipulated what they believed to be the central factor by breaking the circle at any point. We know all this, but we still have lots of malaria.

A pragmatic success rule for PARRY would be that the use of the theory leads to the amelioration or prevention of paranoid conditions. That the earlobe temperature is high in paranoia is a prediction (not yet tested) that startles many who are unable to intuit the remote connection. That a medication desensitizing paranoids to shame would be ameliorative is plausible to many psychiatrists who now understand the theory. That cultures using few shaming techniques for socialization of children will have a low frequency of paranoia is plausible to a few cultural psychiatrists (admittedly they are friends of mine). Lindsay concludes that "the ball is still in Colby's court." What Colby's court needs is a few dozen more players on that side of the net interested in investigative procedures, testing predictions, performing psychophysiological and linguistic experiments, surveying cultures and subcultures, and searching for shame-reducing medications. The problem is too big to be solved (even more or less) by one man alone in the court.

Magaro & Shulman state that they are unclear about the type of paranoia shown by PARRY and refer to Kendler's (1980) review. Kendler suggests renaming paranoia "simple delusional disorder" to distinguish it from schizophrenia and from manic-depressive illness. Kendler's (1980, p. 699) criteria are:

- (1) Onset of illness was before age 60;
- (2) Non-bizarre

delusions of any type (i.e., persecutory, grandiose, somatic, jealousy, etc.) and/or persistent, pervasive ideas of reference were present. These delusions were usually of a fairly systematized nature; (3) Aforementioned symptoms were present for a minimum of two weeks; (4) Persistent hallucinations of any kind were absent; (5) A full affective syndrome, either depressed or manic, was absent when patient was delusional; (6) There were no symptoms suggestive of schizophrenia including prominent thought disorder, inappropriate affect, patently bizarre delusions, and Schneiderian symptoms; (7) Symptoms suggestive of an acute or chronic organic brain syndrome were absent.

Since PARRY meets each and every one of these syndromic criteria, it is puzzling that Magaro & Shulman are puzzled by the type of paranoia under consideration. The model exhibits far more than the single sign of persecutory delusions. Five additional sets of properties are described in section 2.1 of the target article. In many interviews, the persecutory delusions do not even arise, yet the paranoid mode is obvious to expert judges. Notice that the diagnosis is also based on what the model does *not* exhibit (items 4-7). The holes are a semantically central property in identifying Swiss cheese.

We do not claim that shame-humiliation is the "sole" source of persecutory delusions. The careful reader of the proposed lawlike statement discussed midway through section 2.1 will note that we are describing *sufficient* conditions, not *necessary* conditions. It is explicitly stated that the door is open to alternative antecedents and alternative internal mechanisms. The problem of alternative theories is discussed further in our reply to Miller's commentary.

By validity Magaro & Shulman seem to mean construct validity, a type of stability across different situations. PARRY expects the interviewer to be a psychiatrist asking psychiatric questions and not someone interviewing it for a job. Since it is available to everyone on the SUMEX system, most interviewers of the model have been laymen, and, even then, paranoid behavior certainly is evoked whether the interviewer realizes it or not. To answer the question about diminishing the paranoid mode without simply waiting for time decay: There is a way to accomplish this in an interview, but it is hard to find without practice. We saved and studied over 50,000 interviews with the model. (Many outside of AI do not realize how much tedious dog work is involved in data collection and examination as a research resource in implementing theoretical models.) In all these interviews, only once did a first-time interviewer evoke paranoid responses and then guide the subsequent course of the interview in such a way that the paranoia diminished, and the model returned to its nonparanoid mode. The interviewer was an experimental neurophysiologist with no experience in psychiatry!

Maher believes that our interview tests, designed to test the model, tested the theory. Indistinguishability tests at the input/output level do not test the theory; they test the goodness of the *imitation*, regardless of how the imitation was produced. The explanatory

theory of a model can be a good one but yield a poor imitation because of implementation problems. Alternative theories when modeled can produce equally good imitations of performance.

**Maher's** concern about which of the components of the model are essential constitutes the problem of sensitivity analysis, familiar to simulation workers from the earliest days. To get the model going, one postulates and implements a set of factors believed to be responsible. Then one tinkers with the model, removing factors singly or in combination to see whether its behavior is significantly different. Sometimes it turns out that a factor believed to be central has little effect in the overall processing. Sensitivity analysis is an extremely laborious process, encumbered by the great number of combinations of factors that can be manipulated. We conducted a few such experiments with early versions of the model. For example, if the model's output responses are randomly selected, judges do not consider it seriously as a patient. The output must make some sense in response to the input for an interview to be judged as authentic. All sorts of other sensitivity experiments could be conducted, as **Maher** suggests, given time, money, and manpower.

Model versions were tested against many hospitalized paranoid patients, not just the one referred to in the target article (see **Colby** 1975). We also had psychiatrists interview hospitalized alcoholics and depressives over the teletype. None of them was judged to be paranoid. Paranoids are simply very different from other kinds of patients, and it is not a difficult diagnosis to make from the purely verbal data of an interview.

The problem of theory validation or verification arises again in **Miller's** commentary. As **Lindsay** mentioned, the behavioral sciences lack decidability criteria for successful theories. Most have no pragmatic efficacy criteria that can be fed back into the theory. Also, many behavioral scientists are hung up on some notion that theories must be verified, certified, and guaranteed as "true" in spite of the fact that historians of science, philosophers of science, and even leading theoretical physicists have repeatedly instructed us that no theory in science has ever been conclusively proven to be true. "Absolute truth" is a metaphysical, not a scientific, category. Theories are revisable conjectures having evidential support and are useful in organizing, correlating and enlarging our experiences, making them more coherent by unifying them under the aegis of a connecting pattern or invariant. Popperians claim that one cannot prove a theory, but that one can disprove it. But this would mean establishing the truth of the falsifying hypotheses, which is similarly impossible. If predicted observations are used to disprove, one can always question (a) the accuracy of the observations, (b) their underlying assumptions (other theories), (c) the necessary other-things-being-equal clauses, and (d) even the relevance of the particular observation. Between theories and observations lie correspondence rules or epistemic correlations that bridge, or make relevant, the theoretical and observational concepts. One can claim the observations are misclassified in relation to the theory. Since no one can say what truth

is, we should abandon it as a criterion for theories. When a scientist speaks of a "true" theory, he simply means it has demonstrable utility in making sense of, and increasing the coherence of, our experience.

Model explanations, embodying theories, perform a multiplicity of valuable functions. The value of our explanation of paranoia at the moment is that it provides a degree of intelligible unification, and one can build on it; that is, new experiences can proliferate. **Miller** asks, "Does the model make sense out of a seemingly incoherent set of facts?" The theory unifies such diverse facts as that the birth of a deformed child, or a false arrest, or an upbraiding by a teacher in front of a class, or a severely crippling accident, or **Wernicke's** aphasia, are situations known to be precipitants of paranoid conditions. The known data are derivable from the theory. The theory relates these unrelated-to-common-sense situations through an underlying pattern of shame-induced distress that becomes partially relieved by strategies of blaming others. No other rival theory at present systematically accounts for these diverse paranoia-precipitating conditions by organizing them under a single pattern. Some other current theories, such as the homosexual theory, are included under ours as special cases (**Colby** 1977a).

This does not imply other novel theories could not achieve such a unification. We do not claim ours is the *only* explanation for paranoid phenomena. Many antecedent conditions can lead to the same final condition. There are many alternative, and not incompatible, metabolic pathways by which what we eat becomes glycogen stored in the liver. In protein synthesis, six different codons are now known to code for the single amino acid arginine. Not all six were discovered at once. The discovery of the second codon did not rule out the tenability of the first as a mechanism. Similarly, we should expect that further sets of mechanisms, in addition to ours, will be found in paranoid conditions. As **Miller** sensibly points out, different pathogeneses of an illness would suggest different treatments. Different determining mechanisms could even lead to a different taxonomy of paranoia containing, for example, a shame-humiliation kind, and a kind X, where X constitutes an alternative set of generative mechanisms. One starts with a preliminary classification of kinds. A new theory can result in a rearrangement of this taxonomy because it tells us which kinds are more natural than the ones we started with. They are more natural in the sense that they are no longer based only on appearance but on an underlying property (the internal mechanisms) as well. Individuals may appear similar but become discriminated on the basis of underlying properties rather than on the basis of manifest properties generated by them.

Methodological issues are also raised by **Moor**, who feels there is potentially more to AI than cargo cult science. He distinguishes between programs, models, and theories. In the actual everyday work of model construction, we speak in Plattdeutsch about the theory, the model, and the implementation of the model, but in the Hochdeutsch of the literature, we are inclined to speak only of theory and model. **Moor** is

quite right in correcting the common misconception that a listing of the implementation program represents the model. The computer model is what runs at run time and involves the entire computational system. The mention is brief, but he makes the extremely significant point that a computer model can be a model of many theories, just as a given theory can have many computer models. Modelers themselves often overlook this far-reaching principle.

Independent testing of semantically central hypotheses of the theory and independent data are crucial if AI as theoretical psychology is to become more than a cargo cult science going through the right motions but not yielding new knowledge valuable for the enhancement of people's lives. If AI remains entirely bound to its computational instrument, it becomes philosophy-of-mind engineering. Theories and models composed of a large number of hypotheses and auxiliary assumptions have a large number of computable consequences well beyond those derivable in their own heads by the theoreticians. Moor suggests manipulating the model in a variety of ways to see whether unexpected output results. This was precisely our long-range strategy once we had a satisfactory model. Unfortunately, the plan was interrupted by defunding. In the little time we had, we noted a few instances of what we thought was a new, or at least unnoticed, property of paranoid behavior characterizable as on-the-spot delusional boasting. Whether this is so in the case of actual paranoids remains to be seen. In personal communications we have alerted our clinical colleagues to this sort of creative improvisation. Moor thus pinpoints the main justification for constructing and experimenting with complex models of psychopathology. One can tamper and tinker with them without harming anyone. One can discover ways of making the pathological worse as well as better. It is unethical to probe already suffering patients in ways that one can probe computer models, at least until a Computer Subject Protection Committee is formed.

Oatley emphasizes a point that should have been made clearer in the target article. He is right that the strategies of the paranoid mode are unconscious in the Helmholtz sense. A person is consciously aware of only certain products of the underlying mental processes. In computational jargon, these transconscious processes run in compiled versions, inaccessible to higher-language influence except through interpreted versions. Oatley is also perspicacious in noting that it is not just a shame parameter that "drives" the model. The heart of the matter, if one must point at one, lies in the set of self-defining beliefs whose specific meaning content is that the self is defective. These particular beliefs must be activated by input or by autonomous internal processing to raise the value of the shame parameter. Shame then activates distress, which instigates the shame-reducing strategies. Shame is an affect connected to standards or ideals embraced by a person defining himself as worthy. Since there is great variability in the ideas embraced (believed), a great variety of situations can provoke, or fail to provoke, shame. Some people feel ashamed when ridiculed for their dress, others do not. People in some societies are

ashamed to be seen eating, but not defecating. The most fundamental restructuring that can take place in the paranoid mode is in the revision of the core self-defectiveness beliefs or in diminishing repeated activation of them. One of our treatment recommendations, now adopted by some clinicians at UCLA, is to try to help the patient emancipate himself from repeated exposures to humiliating defeats, such as may be occurring in a job or school or sexual situation.

Reducing repeated activation by managing the environment is easier than gaining access to compiled defectiveness beliefs. Products of underlying mental activity appear in our consciousness in the linguistic form of inner speech, audible to the "mind's ear." To fuse Bertrand Russell's distinction between knowledge by acquaintance and knowledge by description, we have knowledge of what we think by acquaintance with descriptions of what we think formulated in natural language, the high-level programming language we use in talking to ourselves. In these inner monologues, we can change our internal representations, given a description of them to work on. (Note: the descriptions are not the representations themselves.) We can restructure our programs, to some extent, using the instructions of the high-level code of natural language. But to conduct this sort of reprogramming, how do we gain access to compiled programs running transconsciously and automatically? In theory, we change some of these compiled programs by rewriting the interpreted version and then recompiling it.

In the "talking" therapies of psychiatry, one tries to restructure some of the patient's representations. In the case of paranoia, this method has not been effective. Perhaps it has been misapplied. No one knows quite how to go about accessing sensitive representations or whether the resultant "understanding" would do any good. As Moor indicated, one way to explore the problem would be to have a good model of paranoia and then experiment with it in ways that one cannot with actual patients. But first one must have a good model.

According to Pinkava, contemporary psychology still suffers from specters of the past - motivational concepts and behaviorism. He proposes we move to the "greener pastures" of the "theory of self-regulating systems," but it is not clear to me whose theory he has in mind and how this helps. As a model, PARRY is a motivational system that shows types of input/output behavior and is self-regulating in the sense of its nonuniqueness of dependence of output on input. Pinkava is right that it is not an ontogenetic explanation of how paranoid conditions originated in the first place; it attempts to explain only what is happening *now* in a particular type of interview. The explanation is pathogenetic only.

Pinkava would like to see a program generate paranoid behavior through a flaw in the program logic. What sort of logic? There is only one way to demonstrate this mechanism, and that is to construct and run such a program. I would not know how to do it myself; hence I can cheerfully pass the task on to others, among them Pinkava. He believes that models such as ours

could never show behavior unexpected by someone familiar with the program. I can assure him that PARRY is full of surprises, not only because of the large number of combinatorial possibilities mentioned above, but also because the model has truly independent variables, namely, the interviewer, what he says, when and how he says it, and how he responds to what is said to him.

I fail to see how our approach is necessarily a "blend of irreconcilable ideas: a mentalistic conception . . . tested in a behaviorist way." Mental predicates such as belief, affect, and the like are theoretical entities. Their nature and function are tested by observation of the behavior (especially linguistic) of people to whom we ascribe these predicates. How else should we do it? Pinkava's argument here amounts to little more than asserting that because mental concepts and behavioral concepts have been around for a long time, there is something incorrect about them. The concept of gravitational force is quite old, quite useful, and still quite fruitful.

For some opaque reason, Reid & Riedler believe we are Chomskyites working on syntactic structures and transformational grammar. We must eschew the honor of these misapplied buzz terms. Reid & Riedler allude to a "paranoid whole." A "whole" is some sort of total system whose components are related in a structure. If we decompose the whole into its components, we lose the structure, the way in which the components are linked or bonded to one another. If we lose the structure, we cannot explain the gestalt or wholeness properties at all because they are a function of the structure and relations of the components, not just the itemized components by themselves. Our theory and model do not account for all the phenomena of paranoid conditions, if that is what is meant by "whole." They are limited to central and defining properties. If anyone wants to explain more properties, he must state what they are and how they are to be accounted for. Thus more and more about paranoia will become understood.

Reid & Riedler object that our efforts would burden the clinician by adding an extra process to his decision making in moving from clinical observations to diagnostic formulations. I plead guilty to being a baneful influence and trying to make clinicians think. They, or someone, should theorize about why manifest properties characterize one taxonomic category rather than another. Perhaps there is something more abstract underlying the manifestations. If that something involves shame-induced distress, then rational treatment recommendations and procedures would involve reducing this type of distress. If psychiatrists are not to think at all, but use only "mature clinical judgment" (the oracle speaks!), then their patients will be unthinkingly treated.

Like so many others, Reid & Riedler show a lamentable lack of discrimination between theoretical model and facsimile, confusing a look-alike model airplane with the abstract principles of aerodynamic lift. They want a clone or duplicate, not an explanatory model. Their particular clinical clone is a bulk replica in the form of a total human mind. At first I thought I would

relish the opportunity of formulating an impaling harpoon to counter the needless needle that our work is "sophomoric." But then I shruggingly decided - ah, the hell with it!

Revlin starts out by claiming that criterial behavior, such as that found in Turing-like tests, is not quantifiable, but, he later says, "counts of accusatory statements" can be made. Interview data are as quantifiable as any other. One selects the scale, the unit, and the rule by which numbers are assigned to observations, just as in any other measurement technique.

Regarding "truthfulness" in evaluation, in the experiment reported in the target article (Heiser et al. 1980), we did in fact inform the judges that the interviews they conducted might involve a computer simulation of a patient. In other experiments, using transcripts sent to judges, both computer scientists and psychiatrists, we informed them that one of the two interviews was with a computer simulation. The "no-significant-difference" hypothesis still held up. People who have not participated in such tests do not quite understand or do not believe, or do not want to believe, how difficult it is to make the correct identification. "Maybe gullible psychiatrists or computer scientists can be fooled by this sleight of mind, but not *me*."

Sloman, adopting a deep-shallow metaphor, wants a more complex model that includes "physiological processes." At what level should models of mental activity be sealed off, above and below, from other relevant knowledge? A model builder must decide on what to include as significant and what to exclude, not as nonexistent, but as negligible for his purposes. Sloman's slogan is "only when we have a deep understanding of . . . [the] 'normal' can we hope to understand . . . [the] pathological." Without unpacking the deep-shallow metaphor, we can note that there are dozens of examples in medicine in which we understand the pathological quite well and do not understand the normal at all. For instance, we understand the pathogenesis of many illnesses, but no one yet understands what it means to be healthy, unless it is defined only as the absence of illness.

In Sloman's ideal model, "feeling afraid involves a complex cognitive state involving beliefs about what is likely to happen and desires that it should not happen." If this is what is meant by "deep," this is exactly how PARRY operates. It becomes afraid when it believes that the interviewer may be linked to the Mafia and therefore will harm it; it desires not to be harmed, and it takes steps to avoid the anticipated harm. If these interpretation-action patterns become implemented in Sloman's planned model, then knowledge in AI *does* accumulate in spite of what snappish critics have often claimed.

When Sloman has his model of the normal up and running (it must be more than an armchair concoction of ingredients), he will be faced with the inescapable taxonomic question, What is your definition or criterion of "normal"? The category "normal" is too vague to qualify as a natural kind that can serve as an explanandum. This lack of initial taxonomic "consensibility" is one of the strongest criticisms leveled at the behavioral sciences. The advantage of "paranoia" as a

natural kind is that it gains a high concordance of agreement. Notice that even among our commentators it does not gain 100% agreement. The fabric of the world is probabilistic, and there is room for disagreement arising from inherent variation in the phenomena, measurement errors, and alternative interpretation. Until there is a starting workable preliminary classification of "normal" kinds of mental functions, the explanatory models of normal processes will have a hard time gaining acceptance. They lack the first step of taxonomic agreement as to what it is that we are trying to explain.

Swanson raises the issue of a difference between appearing paranoid and being paranoid. If one trained an actor to appear paranoid, he might succeed in getting judges to believe he was actually paranoid. How was the actor trained? He cannot simply repeat over and over that the Mafia is after him. To succeed in the deception, he must be using a set of rules or general principles applicable to diverse input situations to produce the appropriate behavior. What rules and how many rules must he know consciously? Can he remember 200 rules? How correctly must he apply them to appear paranoid? His success would then not *prove* the rules, it would demonstrate their sufficiency and feasibility. It would make it plausible that these might be the rules unconsciously guiding the paranoid. Being paranoid would then imply operating under these rules. Our model applies rules to exhibit paranoid behavior. It is a demonstration of our model representation, a *Darstellung* of a *Vorstellung*. The demonstration adds a plus to the notion that the theory is not an idle entry in the lists of theories, but a candidate for acceptability as a partial solution to the problem of paranoia.

Models that interact with us in natural language tend to bring out playfulness. Most of the interviewers of PARRY over the nationwide SUMEX network have not been mental health professionals but AI workers and graduate students who enjoy playing all sorts of games with the model. A favorite one is simply to swear at it endlessly. This spirit of playfulness is exemplified in the comments of Teller & Dahl, whose cute scenario merits dramatic as well as scientific consideration. I will restrict myself to the latter as much as possible.

Teller & Dahl point out that paranoia is easily diagnosed, with the mention of the Mafia being a giveaway. Many of our expert judges diagnose the model as paranoid without the delusions regarding the Mafia appearing in the interview. The hint of delusion and the secondary phenomena such as hypersensitivity, self-reference, and unjustified hostility are often sufficient for a judge to formulate a diagnosis of paranoia.

Why Teller & Dahl feel the theory and model are at a dead end is cloudy to me. As already indicated, all sorts of science-extending observational and experimental work spaces fan out from the implications of the theory. In this way, theories can expand and enlarge our experience. The question for the future is whether anyone will work in these spaces, especially now with the large cutbacks in funded research.

In the denouement of the scenario, D. turns out to be an excellent simulation of Dahl, a description of whom

is quoted from a *New Yorker* magazine hype. But if the simulation is that good, how do we know the *New Yorker* reporter was not actually interviewing D. rather than Dahl?

## References

- Abelson, R. P. (1968) Computer simulation of social behavior. In: *Handbook of social psychology*, ed. G. Lindzey & E. Aronson, vol. 2. Reading, Mass.: Addison-Wesley. [RR]
- (1973) The structure of belief systems. In *Computer models of thought and language*, ed. R. C. Schank & K. M. Colby. San Francisco: W. H. Freeman. [JGC, KMCa]
- Agassi, J. (1976) Review of Weizenbaum (1976). *Technology and Culture* 17:13-6. [JA]
- (1977) *Towards a rational philosophical anthropology*. The Hague: Martinus Nijhoff. [JA]
- Alston, W. P. (1977) Self-intervention and the structure of motivation. In: *The self: psychological and philosophical issues*, ed. T. Mischel. Oxford: Blackwell. [KMCa]
- Angrist, B. M. & Gershon, S. (1970) The phenomenology of experimentally induced amphetamine psychosis: Preliminary observations. *Biological Psychiatry* 2:95-107. [DWS]
- Bobrow, D. G. & Winograd, T. (1977) An overview of KRL, a knowledge representation language. *Cognitive Science* 1:3-46. [KMCa]
- Bunge, M. (1973) *Method, model and matter*. Dordrecht: D. Reidel. [KMCa]
- Cameron, N. (1959) Paranoid conditions and paranoia. In: *American handbook of psychiatry*, vol. 1, ed. S. Arieti. New York: Basic Books. [KMCa]
- Carbonell, J. G. (1979) Subjective understanding: Computer models of belief systems. Ph.D. dissertation, Yale University. [JGC]
- (1981) POLITICS: An experiment in subjective understanding and integrated reasoning. In: *Inside computer understanding: Five programs plus miniatures*, ed. R. C. Schank & C. K. Riesbeck. Hillsdale, N.J.: Erlbaum. [JGC]
- Carbonell, J. G. & Hayes, P. J. (1981). Dynamic strategy selection in flexible parsing. *Proceedings of the 19th Meeting of the Association for Computational Linguistics*. [JGC]
- Cerf, V. (1973) Parry encounters the doctor. *Datamation* 19:62-64. [VT]
- Chomsky, N. (1965) *Aspects of a theory of syntax*. Cambridge, Mass.: M.I.T. Press. [JGC]
- Churchland, P. M. (1979) *Scientific realism and the plasticity of mind*. Cambridge: Cambridge University Press. [KMCa]
- Clippinger, J. (1977) *Meaning and discourse: A computer model of psychoanalytic speech and cognition*. Baltimore: The Johns Hopkins University Press. [VT]
- Colby, K. M. (1963) Computer simulation of a neurotic process. In: *Computer simulation of personality: Frontier of psychological research*, ed. S. S. Tomkins & S. Messick. New York: Wiley. [KMCa, RR]
- (1973) Simulation of belief systems. In: *Computer models of thought and language*, ed. R. Schank & K. Colby. San Francisco: W. H. Freeman. [RR]
- (1975) *Artificial paranoia: Computer simulation of paranoid processes*. Elmsford, N.Y.: Pergamon Press. [KMCar]
- (1977a) An appraisal of four psychological theories of paranoid phenomena. *Journal of Abnormal Psychology* 86:54-59. [KMCa]
- (1977b) On the way people and models do it. *Perspectives in Biology and Medicine* 21:99-104. [KMCa]
- (1978) Mind models: An overview of current work. *Mathematical Biosciences* 39:159-85. [KMCa]
- Colby, K. M.; Faight, W. S. & Parkison, R. C. (1979) Cognitive therapy of paranoid conditions: Heuristic suggestions based on a computer simulation model. *Cognitive Therapy and Research* 3:55-60. [KMCa]
- Colby, K. M.; Hilf, F. D.; Weber, S. & Kraemer, H. (1972) Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence* 3:199-221. [KMCa, JRM, RR]
- Colby, K. M. & Spar, J. E. (1981, forthcoming) *The core problem in psychiatry: Diagnostic taxonomy*. [KMCar]
- Dennett, D. C. (1969) *Content and consciousness*. New York: Humanities Press. [KMCa, KG]
- (1972) Reply to Arbib and Gunderson. *Journal of Philosophy* (reprinted in Dennett 1978). [KG]
- (1978) *Brainstorms*. Montgomery, Vt.: Bradford Books. [KMCa, KG]

- Descartes, R. (1637) *Discourse on method. The philosophical works of Descartes*, vol. 1, trans. E. Haldane & R. T. Ross. New York; Elsevier 1955. [KMCa, KG]
- DSM-III (1980) *Diagnostic and statistical manual of mental disorders*. 3rd ed. Washington, D.C.: American Psychiatric Association. [KMCr, PAM, VT]
- Faught, W. S. (1978) *Motivation and intentionality in a computer simulation model of paranoia*. Basel: Birkhauser Verlag. [KMCa, PAM]
- Fodor, J. A. (1968) *Psychological explanation: An introduction to the philosophy of psychology*. New York: Random House. [KMCa]
- (1975) *The language of thought*. New York: Thomas Y. Crowell. [KMCa]
- Freedle, R. O. (1972) Language users as fallible information processors. In: *Language comprehension and the acquisition of knowledge*, ed. R. O. Freedle & G. B. Carroll. Washington, D.C.: V. H. Winston. [KMCa]
- Freud, S. (1922) Some neurotic mechanisms in jealousy, paranoia and homosexuality. *Standard Edition* 18:223-34. London: Hogarth Press. [VT]
- Fried, Y. & Agassi, J. (1976) *Paranoia: A study in diagnosis*. Boston Studies in the Philosophy of Science 50. Dordrecht and Boston Reidel. [JA]
- Goffman, E. (1959) *The presentation of self in everyday life*. New York: Doubleday Anchor. [JA]
- Grice, H. P. (1975) Logic and conversation. In: *Syntax and semantics*, vol. 3, ed. P. Cole & J. Morgan. New York: Academic Press. [KMCa, RR]
- Griffith, J. D.; Cavanaugh, J.; Held, J. & Oates, J. A. (1972) Dextroamphetamine: Evaluation of psychomimetic properties in man. *Archives of General Psychiatry* 26:97-100. [DWS]
- Grzegorzczak, A. (1961) *Zarys Logiki Matematycznej* (in Polish) (An outline of mathematical logic). Warsaw: Panstwowe Wydawnictwo Naukowe. [VP]
- Gundersen, K. (1964) The imitation game. *Mind* 73:234-45. [KG]
- (1970) Asymmetries and mind-body perplexities. In: *Minnesota Studies in the Philosophy of Science*, vol. 4, ed. M. Radner & S. Winokur, pp. 273-309. Minneapolis: University of Minnesota Press. [KG]
- (1971) *Mentality and machines*. New York: Doubleday and Company. A revised and expanded edition of this book is forthcoming from University of Minnesota Press in 1982. [KMCr, KG]
- (1972) *Content and consciousness* and the mind-body problem. *Journal of Philosophy* 69:591-604. [KG]
- (1974) The Texture of Mentality. In: *Wisdom - Twelve essays*, ed. R. Bambrough, pp. 173-95. Oxford: Blackwell. [KG]
- Hantler, S. & Kochen, M. (1973) ASP: A program using stored hypotheses to select actions. *Journal of Cybernetics* 3:1-12. [MK]
- Harre, R. (1970) *The principles of scientific thinking*. Chicago: University of Chicago Press. [KMCa]
- (1977) Science as representation. *Philosophy of Science* 44:146-58. [KMCa]
- Harre, R. & Secord, P. F. (1972) *The explanation of social behavior*. Oxford: Blackwell. [KMCa]
- Harris, Z. S. (1962) *String analysis of sentence structure*. New York: Mouton and Co. [KMCa]
- Hayes-Roth, P.; Fox, M.; Gill, G. & Mostow, D. J. (1976) Semantics and pragmatics in the Hearsay II speech understanding system. In: *Speech understanding systems: Summary of results of the five-year research effort*, ed. R. Reddy. Department of Computer Science, Carnegie-Mellon University. [KMCa]
- Heiser, J. F.; Colby, K. M.; Faught, W. S. & Parkison, R. C. (1980). Can psychiatrists distinguish a computer simulation of paranoia from the real thing? The limitations of Turing-like tests as measures of the adequacy of simulation. *Journal of Psychiatric Research* 15:149-62. [KMCr]
- Hendrix, G. G. (1977) The LIFER manual: A guide to building practical natural language interfaces. Stanford Research Institute AI Center, Technical note 138. [KMCr]
- Hendrix, G. G., Sacerdoti, E. D. & Slocum, J. (1976) Developing a natural language interface to complex data. Tech. report Artificial Intelligence Center, SRI International. [JGC]
- Hilf, F. D. (1972) Non non-verbal communication and psychiatric research. *Archives of General Psychiatry* 27:631-35. [KMCa]
- Hofstadter, D. R. (1979) *Gödel, Escher, Bach: An eternal golden braid*. New York: Basic Books. [KMCr, VT]
- (1981) A coffeehouse conversation on the Turing test to determine if a machine can think. *Scientific American*, 244:15-36. [KMCr]
- Holly, W. J. (1975) Program-resistant aspects of knowing-that and knowing-how. Doctoral dissertation, University of California at Irvine. [KG]
- Izard, C. E. (1971) *The face of emotion*. New York: Appleton-Century-Crofts. [CEI]
- (1972) *Patterns of emotions: A new analysis of anxiety and depression*. New York: Academic Press. [CEI]
- (1977) *Human emotions*. New York: Plenum Press. [CEI]
- Kaplan, R. (1973) A general syntactic processor. In: *Natural language processing*, ed. R. Rustin. New York: Algorithmic Press. [KMCa]
- Kay, M. (1973) The mind system. In: *Natural language processing*, ed. R. Rustin. New York: Algorithmic Press. [KMCa]
- Kendler, K. S. (1980) The nosologic validity of paranoia (simple delusional disorder). *Archives of General Psychiatry* 37:699-706. [KCR, PAM]
- Kochen, M. (1960a) Cognitive mechanism. IBM Report, RAP 16. [MK]
- (1960b) Experimental study of hypothesis formation by computer. In: *Information theory*, ed. C. Cherry. London: Butterworth. [MK]
- (1962) Adaptive mechanisms in digital concept-processing. In: *Discrete adaptive processes: Proceedings of the Joint Automatic Control Conference*, American Institute of Electrical Engineers. [MK]
- (1971) Cognitive learning processes: An explication. In: *Artificial intelligence and heuristic programming*, ed. N. V. Findler & B. Meltzer. Edinburgh: Edinburgh University Press. [MK]
- (1974) Representations and algorithms for cognitive learning. *Artificial Intelligence* 5:199-216. [MK]
- Kochen, M. & Galanter, E. (1958) The acquisition and utilization of information in problem-solving and thinking. *Information and Control* 1:267-88. [MK]
- Kochen, M.; MacKay, D. M.; Moran, M. E.; Scriven, M. & Uhr, L. (1967) Computers and comprehension. RAND memorandum, RM-4065-PR. In: *The growth of knowledge*, ed. M. Kochen. New York: Wiley. [MK]
- Kochen, M. & Stark, J. (1978) Representation and formation of hypotheses in learning programs. *Proceedings of the International Conference on Cybernetics and Society*, pp. 69-74. Institute of Electrical and Electronic Engineers, Tokyo. [MK]
- Kuhn, T. S. (1977) *The essential tension: Selected studies in scientific tradition and change*. Chicago: University of Chicago Press. [JGC]
- Lazarus, R. S. (1966) *Psychological stress and the coping process*. New York: McGraw-Hill. [JA]
- Leatherdale, W. H. (1974) *The role of analogy, model and metaphor in science*. New York: Elsevier. [KMCa]
- McCorduck, P. (1979) *Machines who think*. San Francisco: W. H. Freeman. [RR]
- Magaro, P. A. (1980) *Cognition in schizophrenia and paranoia: The integration of cognitive processes*. Hillsdale, N. J.: Erlbaum. [PAM]
- Malcolm, J. (1980) The impossible profession, part 2. *New Yorker* 41:54-152 (December 1). [VT]
- Meehan, J.R. (1976) The metanovel: Writing stories by computer. Ph.D. dissertation, Yale University. [JGC]
- Miller, G. A. & Johnson-Laird, P. N. (1976) *Language and perception*. Cambridge, Mass.: Harvard University Press. [KMCa]
- Miller, J. R. & Kintsch, W. (1980) Readability and recall of short passages: A theoretical analysis. *Journal of Experimental Psychology* 6:335-54. [JRM]
- Minsky, M. (1963) Steps towards artificial intelligence. In: *Computers and thought*, ed. E. Feigenbaum & J. Feldman. New York: McGraw-Hill. [KG]
- Moor, J. H. (1976) An analysis of the Turing test. *Philosophical Studies* 30:249-57. [KMCr, JHM]
- (1978) Three Myths of Computer Science. *British Journal for the Philosophy of Science* 29:213-22. [JHM]
- Moore, J. & Newell, A. (1973) How can MERLIN understand? In: *Knowledge and cognition*, ed. L. W. Gregg. Erlbaum. [KMCa]
- Nagel, T. (1979) *Mortal questions*. Cambridge: Cambridge University Press. [KG]
- Nash-Webber, B. L. (1977) Anaphora: A cross disciplinary survey. TR 31. Urbana, Ill.: The Center for the Study of Reading. [KMCa]
- Newell, A. & Simon, H. A. (1972) *Human problem solving*. Englewood Cliffs: Prentice-Hall. [KMCa]
- Norman, D. A. (1980) Twelve issues for cognitive science. In: *Perspectives on cognitive science: Talks from the La Jolla conference*, ed. D. A. Norman. Hillsdale, N.J.: Erlbaum. [CEI]
- Panofsky, E. (1954) *Galileo as a critic of the arts*. The Hague: Martinus Nijhoff. [KMCa]
- Parkison, R. C. (1980) An effective computational approach to the comprehension of purposeful English dialogue. Ph.D. dissertation, Department of Computer Science, Stanford University. [KMCa, PAM]
- Pinkava, V. Z. J. (1980) ENIGMA: A computer program you can talk to. Unpublished 3rd year undergraduate project, Oxford University. [VP]
- Platt, J. R. (1964) Strong inference. *Science* 146:347-53. [RL]
- Potkin, S. G.; Karoum, F.; Chuang, L.-W.; Cannon-Spoor, H. E.; Phillips, I.; and Wyatt, R. J. (1979) Phenylethylamine in paranoid chronic schizophrenia. *Science* 206:470-71. [DWS]
- Pylyshyn, Z. (1980) Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences* 3:111-69. [KMCa]

## References/Colby: Modeling a paranoid mind

- Rescher, N. (1970) *Scientific explanation*. New York: Free Press. [KMCa]
- Retterstol, N. (1966) *Paranoid and paranoid psychoses*. Springfield, Ill.: Charles C. Thomas. [KMCr]
- Riesbeck, C. & Schank, R.C. (1976) *Comprehension by computer: Expectation-based analysis of sentences in context*. Tech. report 78, Computer Science Department, Yale University. [JGC]
- Rosehan, D. L. (1973) On being sane in insane places. *Science* 179:250-58. [JA, BAM, DWS]
- (1975) The contextual nature of psychiatric Diagnosis. *Journal of Abnormal Psychology*, 84:462-74. [JA]
- Ryle, G. (1949). *Concept of mind*. New York: Barnes and Nobel Books. [KG]
- Schank, R. C. (1975) *Conceptual information processing*. Amsterdam: North-Holland. [JGC]
- Schank, R. C. & Abelson, R. P. (1977) *Scripts, plans, goals and understanding*. Hillsdale, N.J.: Erlbaum. [KMCa, JRM]
- Schank, R. C. & Colby, K. M., eds. (1973) *Computer models of thought and language*. San Francisco: W. H. Freeman. [KMCa]
- Searle, J. R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3:417-57. [RPA, KMCa, RL, JRM]
- Simon, H. (1969) Motivational and emotional controls of cognition. *Psychological Review* 74:29-39. [KMCa, KG]
- (1980) Cognitive science: The newest science of the artificial. *Cognitive Science* 4:33-46. [KMCa]
- Sloman, A. (1978) *The computer revolution in philosophy: Philosophy, science and models of mind*. Atlantic Highlands, N. J.: Humanities Press. [KMCa]
- (1981) Why robots will have emotions. Paper presented at the International Joint Conference on Artificial Intelligence, Vancouver, Canada, August, 1981. [RPA, JGC]
- Spitzer, R. L.; Forman, Y. B. W. & Nee, J. (1979) DSM-III field trials: I. Initial interrater diagnostic reliability. *American Journal of Psychiatry* 136:815-17. [KMCa]
- Swanson, D. W.; Bohnert, P. J. & Smith, J. A. (1970) *The paranoid*. Boston: Little, Brown. [KMCa]
- Szasz, T. S. (1961) *The myth of mental illness*. New York: Hoeber-Harper. [JA]
- (1970) *The manufacture of madness*. New York: Harper. [JA]
- Tomkins, S. S. (1962) *Affect, imagery, consciousness*. Vol. 1, *The positive affects*. New York: Springer. [CEI]
- (1963) *Affect, imagery, consciousness*. Vol. 2, *The negative affects*. New York: Springer. [KMCa, CEI]
- Turing, A. (1950) Computing machinery and intelligence. In: *Computers and thought*, ed. E. A. Feigenbaum & J. Feldman. New York: McGraw-Hill. [KMCa, KG, RR]
- Vigotsky, L. S. (1962) *Thought and language*. New York: Wiley. [JA]
- Wang, H. (1960) Toward mechanical mathematics. *IBM Journal of Research and Development* as cited by M. Minsky (1963) Steps toward artificial intelligence. In: *Computers and thought*, ed. E. A. Feigenbaum & J. Feldman. McGraw-Hill. [KG]
- Weizenbaum, J. (1976) *Computer power and human reason*. San Francisco: W. H. Freeman. [JA, JRM]
- Wilks, Y. (1973) Preference semantics. Stanford Artificial Intelligence Laboratory, Memo AIM 206. [KMCa]
- Winograd, T. (1972) *Understanding natural language*. New York: Academic Press. [KMCa]
- Woods, W., Kaplan, R. & Nash-Webber, B. (1972) The lunar sciences natural language information system: Final report. Tech. report 2378, Bolt Beranek and Newman Report. [JGC]
- Zajonc, R. B. (1980) Feeling and thinking: Preferences need no inferences. *American Psychologist* 35:151-75. [CEI]