Stanford Artificial Intelligence Laboratory
Memo AIM-305

September 1977

Computer Science Department
Report No. STAN-CS-77-633

MOTIVATION AND INTENSIONALITY IN A COMPUTER SIMULATION MODEL

by

William S. Faught

Research sponsored by

National Institutes of Health
and
Advanced Research Projects Agency

COMPUTER SCIENCE DEPARTMENT
Stanford University

Stanford Artificial Intelligence Laboratory
Memo AIM-305

September 1977

Computer Science Department
Report No. STAN-CS-77-633

MOTIVATION AND INTENSIONALITY IN A COMPUTER SIMULATION MODEL

by

William S. Faught

ABSTRACT

This dissertation describes a computer simulation model of paranoia. The model mimics the behavior of a patient participating in a psychiatric interview by answering questions, introducing its own topics, and responding to negatively-valued (e.g., threatening or shame-producing) situations.

The focus of this work is on the motivational mechanisms required to instigate and direct the modelled behavior.

The major components of the model are:

(1) A production system (PS) formalism accounting for the instigation and guidance of behavior as a function of internal (affective) and external (real-world) environmental factors. Each rule in the PS is either an action pattern (AP) or an interpretation pattern (IP). Both may have either affect (emotion) conditions, external variables, or outputs of other patterns as their initial conditions (left-hand sides). The PS activates all rules whose left-hand sides are true, selects the one with the highest affect, and performs the action specified by the right-hand side.

(2) A model of affects (emotions) as an anticipation mechanism based on a small number of basic pain-pleasure factors. Primary activation (raising an affect's strength) occurs when the particular condition for the affect is anticipated (e.g., anticipation of pain for the fear affect). Secondary activation occurs when an internal construct (AP, IP, belief) is used and its associated affect is processed.

(3) A formalism for intensional behavior (directed by internal models) requiring a dual representation of symbol and concept. An intensional object (belief) can be accessed either by sensing it in the environment (concept) or by its name (token). Similarly, an intensional action (intention) can be specified either by its conditions in the immediate environment (concept) or by its name (token).

Issues of intelligence, psychopathological modelling, and artificial intelligence programming are discussed. The paranoid phenomenon is found to be explainable as an extremely skewed use of normal processes. Applications of these constructs are found to be useful in AI programs dealing with error recovery, incompletely specified input data, and natural language specification of tasks to perform.

This thesis was submitted to the Department of Computer Science and the Committee on Graduate Studies of Stanford University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was supported by the National Institutes of Health under Contract PHS NIMH 06645-13 and Advanced Research Projects Agency of the Department of Defense under ARPA Order No. 2494, Contract MDA903-76-C-0206. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Stanford University or any agency of the U. S. Government.

Preface

The work presented in this dissertation is part of an attempt to develop a simulation model of a particular theory of paranoia. The original goal was to simulate the behavior of a psychiatric patient in a first diagnostic interview. For this purpose it was necessary to develop, in addition to the specific embodiment of the theory of paranoia, a natural language interface and a number of behavioral responses to the myriad inputs possible. It soon became obvious that extensive normal mental processing would also have to be modelled. Also, as the theory of paranoia was elaborated and extended, it relied more and more on an underlying base of normal mental processes.

From the exploration of these normal processes, a second goal emerged. This goal is to develop a model of human motivational processes. It stems from the author's wish to understand what the top level execution loop in humans is, and therefore what the top level execution loop in autonomous intelligent computers ought to be. The focal point of this dissertation is the set of these motivational processes.

This dissertation attempts to:

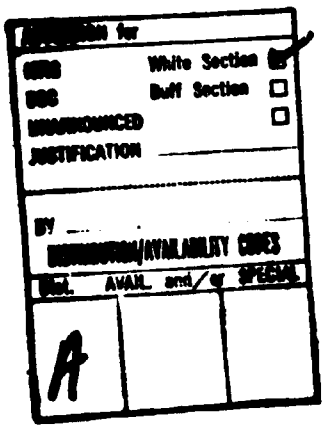
(1) show the necessity for modelling motivational and intensional constructs in models of human behavior (intensional constructs being characterized by internal representation);

(2) outline a simulation model of motivational processes as a base for a model of paranoia; and

(3) discuss the implications of cognitive direction of conative processes for intelligent computer programs.

Most of this research was performed at the Stanford Artificial Intelligence Laboratory and at the Department of Psychiatry at UCLA. I am indebted to their principal sponsors, the National Institute of Mental Health and the Advanced Research Projects Agency, for their support.

A great many people helped me, directly and indirectly, to complete this work. I wish to thank my adviser, Dr. Kenneth Colby, for his questions and suggestions, and for his untiring efforts to make sense of the human mind. I am grateful to Roger Parkison for his analysis and suggestions while my ideas were in the formative stage. I am also indebted to the members of my reading committee, Bruce Buchanan and Terry Winograd, for their helpful comments and suggestions. Finally, I would like to thank all of those who came into contact with my work and helped in the completion of the ideas and manuscript, including David Shaw, Carole Parkison, Erin Colby, Cathy Villa, and Bob Filman.



CONTENTS

Chapter 1. Introduction

The basic question of psychology is: Why do men and animals behave as they do?

--Boden

Section 1.1 Overview

When a person acts, his behavior is guided by several motivational factors. Among these factors are his internal needs (affects or emotions), sensory input from the external environment, and internal mental constructs that represent the environment. To model the motivation processes in a computer, we need a program which combines these motivational factors in such a way that they operate as the sole motivation of the system and they guide processing decisions at the most detailed level.

This paper presents one way to model motivational processes in a simple and consistent structure. We implement the internal processes in a special production system (PS). The PS incorporates rules for two types of processing: sequentially-performed actions and parallel interpretation of situations. The factors that determine rule selection are affects, environmental conditions, and intensional constructs.

We believe that implementing complex theoretical models of human behavior in computer simulations is a reasonable method of exploring those models. The computer program in our work (PARRY3) simulates a paranoid patient. A psychiatrist sits at a teletype and interviews the program as if it were a patient. The psychiatrist's task is to conduct a first diagnostic interview with the model-patient. The program responds by answering questions, introducing its own topics, and dealing with negatively-valued (e.g., threatening or shame-producing) situations.

We had three goals in constructing the simulation model. The first was to develop a model of human motivation mechanisms that could account for the various factors that influence human behavior. Although the task domain is limited to the behavior of a human over a time span of about thirty minutes and the subject matter is limited to a psychiatric interview, the range of potential situations is large enough to provide a wide variety of behavior to explain.

The second goal was to explore a theory of paranoia. Earlier model versions of the same theory contained special processes for the paranoid phenomena. In the current model, we implemented a mechanism capable of generating paranoid behavior as an extension of the "normal" motivational processes. This implementation provides a consistency test for both normal and abnormal processes.

The third goal was to construct a performance model of the theory of paranoia to discover the consequences of the theory. A performance model is useful when cognitive models become so complex that a person cannot determine all of the interactions by hand-simulating the theory. A performance model also helps to insure the theory's completeness by requiring the theory builder to account for all of the behavior within the task to be performed.

The paper is organized as follows.

Chapter 1 describes in general the problem to be solved and the goals of the work. It presents a detailed sample interaction between a person and the model and points out examples of motivational and paranoid processes at work.

Chapter 2 presents an overview of the simulation model. Section 2.1 explains the task of the program and gives examples of human behavior typically occuring in a psychiatric interview. The remaining sections of Chapter 2 cover different parts of the model.

Section 2.2 describes the types of mental events and outer physical behavior that humans can perform and that our system must produce. This section first describes human actions. Sequentially-performed actions are distinguished from parallel recognition or interpretation of situations. Actions can be linked into multiple-step (sequential) actions. They can be interrupted if the initial specifying conditions change. Actions can be subdivided into mechanical (nonpurposive) actions performed in a stimulus-response manner and goal-oriented (intensional) actions. Our model generates this behavior using a special production system (PS) to perform the various types of processing. Production rules may be either Action Patterns to specify actions to be performed or Interpretation Patterns to specify situations to be recognized. Interrupts in the PS are realized by changes in the conflict set in a multiple-step action pattern. The three types of conditions which activate the production rules are affects (emotions), environmental conditions, and intensional constructs (beliefs and intentions).

Section 2.3 describes the minimal set of motivational constructs necessary for the model to respond appropriately to the various situations in a psychiatric interview. This section explores the relation between motivational constructs and intelligent behavior. Six affects are represented in the model. They are activated in two ways: (1) The PS interpreter is responsible for primary activation; (2) the right-hand sides of the production rules contain names of affects for secondary activation. Both activations set the affect to higher values. The affects then decay over time to their unactivated levels. The response of the system to affects is realized in two ways: (1) by the affect levels being sensed in the left-hand sides of rules, and (2) by the selection of an action to perform. The actions are ordered according to their affect response content. The PS selects and performs the action with the highest affect content in its response.

Section 2.4 describes the use of symbolic representations to simplify data about the environment. Intensional symbols are characterized by their naming or description feature. The link between a concept and its name or token is traversable in both directions. Without intensional constructs the system cannot describe nonimmediate reality and therefore can only respond in a stimulus-response (nonpurposive) manner. The representation of these constructs is modelled in a manner consistent with the AP-IP motivational mechanism described earlier. The two types of intensional constructs used in the system are beliefs and intentions.

Section 2.5 gives the details of the implementation of the system. This section describes the interface to the language recognizer and gives examples of the various data files that contain patterns, beliefs, intentions, and affect links to beliefs. The actions that are implemented as LISP programs are contrasted to the mental events produced by the PS model.

Chapter 3 presents other related work on problem solving, interpretation of human actions, simulations of motivation, and belief systems. It analyzes the strengths and limitations of their ability to account for the motivation of human behavior.

Chapter 4 describes a production system (PS) formalism accounting for the instigation and guidance of behavior as a function of internal (affective) and external (real-world) environmental factors. Each rule in the PS is either an action pattern (AP) or an interpretation pattern (IP). Both types of patterns may have affect conditions, external variables, and outputs of other patterns as their initial conditions (left-hand sides). The PS activates all rules whose left-hand sides are true, combines the action elements of the activated patterns (conflict set) into a set of multiple attribute actions, selects the action with the highest affect, and executes the selected action.

Chapter 5 describes a model of affects (emotions) as an anticipation mechanism based on a small number of basic pain-pleasure factors. Primary activation (raising an affect's strength) occurs when the particular condition for the affect is anticipated (e.g., anticipation of pain for the fear affect). Secondary activation occurs when an internal construct (AP, IP, belief) is used and its associated affect is processed.

Chapter 6 presents a formalism for intensional behavior (directed by internal models) requiring a dual representation of name and concept. An intensional object (e.g., a belief) can be accessed either by sensing it in the environment (concept) or by its name. Similarly, an intensional action (e.g., an intention) can be specified either by its conditions in the immediate environment (concept) or by its name.

Chapters 4, 5, and 6 outline three essentially different aspects of motivational and intensional behavior: situation-action rules, affects, and symbolic manipulation. Chapter 7 presents the results of how these three interact in larger issues. The issues: (a) The construction of meta systems to observe, interpret, and correct errors in their corresponding object systems. The meta system is activated whenever the base object system fails to find an appropriate action for a situation or whenever the model is awaiting another input from the interviewer. (b) The construction of self-motivating systems and their requirements. The section gives a summary of the features that a self-motivating or autonomous system should have. (c) The elaboration of the theory of paranoia that we set out to model. We see the paranoid mode as a skewed use of normal processes. Implications for the cognitive therapy for patients suffering the pathology are discussed. The major difficulty with therapy is the problem of changing the shame-producing beliefs that a paranoid patient has without setting off the paranoid mode reaction.

Chapter 8 presents several meta issues and implications for artificial intelligence programs.

Section 1.2 The humiliation theory of paranoia

The theory we are attempting to model, the humiliation theory, is more fully stated by Colby [1975]. (Other common theories are the homeostatic, hostility, and homosexual theories of paranoia [Colby 1977].) This theory assumes that there is a paranoid mode of thought which influences certain types of cognitive processing. In the paranoid mode, a person scans natural language input, and the inferences from that input, looking for evidence implying the self is inadequate or defective. Upon finding such evidence, an attempt is made at simulating acknowledgement of this inadequacy. If the person accepts or acknowledges the belief in inadequacy as true, humiliation results. Humiliation represents a negatively-valued affect state signifying a self who is unacceptable in the eyes of the self as well as in the eyes of others. The detection of impending humiliation in the simulation serves as a warning not to execute the acknowledging procedure. Instead the person attempts an alternative simulation in which wrongdoing is attributed to others. Since no warning signal of humiliation results, the person executes the procedure for blaming others. The outcome of this alternate strategy is: (1) to repudiate the assertion that the self is to blame for inadequacy, and (2) to ascribe blame to other human agents. This transfer of blame is reflected in the interpersonal behavior of the paranoid person.

The assumptions in the theory concerning the role of defensive mechanisms in coping with the environment owe much to the theoretical groundwork originally laid down by Freud [Boden, 1974]. For instance, the paranoid mechanism is an example of a strategy used in conflicts to protect part of the self, in this case to protect the self from humiliation. Some of the strategies may distort or hide reality so that the result in consciousness is different from the result that would have occurred without

the strategy. Further, the motivating principle behind such strategies is not rational thought but an irrational, or perhaps nonrational, desire to forestall or reduce a negative affect such as humiliation. The goal in this work is to develop a simple, consistent mechanism that can generate paranoid behavior. In order to attain this goal, we must explain nonrational motivation in terms of affects, environmental conditions, and intensional constructs. Our model of affects and their phenomenology derive from differential emotion theory as proposed by Tomkins [1962] and Izard [1971]. The role of affects in human action is based on work by Mischel [1969].

Section 1.3 Psychopathology

There is one particular set of phenomena which exhibits itself in human behavior that merits attention when studying motivational processes. That set of phenomena is known as psychopathological behavior.

Given that humans are intelligent, that humans have the ability to adapt, one might ask why humans sometimes fail to adapt successfully. The prominent reasons, from an information processing viewpoint, are: (1) a failure in a human's hardware or physiology; (2) a failure in the individual's ontogenetic development, such that part of the human's training may have been omitted or incorrect; (3) a failure to recognize an internal need and the need not being satisfied; (4) a failure to recognize a new external change in the environment, at least until sufficient harm has come to the individual. But even persons whose development seems to have been adequate, who seem to have all of their physiological mechanisms in working order, and who can adapt successfully in certain kinds of situations, fail to adapt. This failure occurs even after the new situation has been in effect a sufficient amount of time, giving the individual an opportunity to adapt. The person fails to adapt and continually performs behavior to his ultimate disadvantage.

We define psychopathological behavior in humans as behavior which humans maladaptively and repeatedly use to their own detriment in particular situations. One problem is defining what a person's detriment is. A criminal, if he is caught and punished, has performed behavior that is to his disadvantage. A person can unknowingly perform behavior that is to his disadvantage. But we will attempt to restrict our interest to a standard, well-defined and well-accepted psychopathology, the paranoid phenomena.

There are numerous reasons for studying pathological behavior in the context of exploring motivation. First, pathological behavior represents human behavior at some kind of extreme. If it is explained by a paradigm we propose for the remainder of human behavior, it then gives us an opportunity to observe some of the boundary conditions for normal behavior. Second, pathological behavior involves almost all aspects of normal human psychological processing. Third, it seems that this behavior is called upon by the individual to solve the most crucial problems of his existence. Fourth, some abnormal states (e.g., paranoia) are not due to hardware failures (according to our theory) but must be explained in terms of the normal mechanisms being exposed to special environmental conditions which skew them into their stereotypical behavior. Finally, it seems that the times when the individual is least efficient at coping with the environment, and the times when he is suffering most, are associated with this aberrant behavior.

Two examples of pathological behavior are stage fright and paranoia. Stage fright, in its mildest form, is the natural fear of failure to perform in frunt of a large group of people. In a mild form, stage fright can actually be beneficial by raising a person's commitment, and thus his resources such as adrenaline, for performance. But in its extreme form, stage fright can be paralyzing, and therefore actually be detrimental to the benefit of the individual. Paranoia, the example we will discuss

extensively in this work, can be paralyzing in its extreme when the person blames others for any discomfort. The existence of these psychopathological behaviors caused earlier psychologists to hypothesize the "disruptive" or "disorganizational" properties of emotion, since behavior with emotional content seemed to lead the organism to its disadvantage. The explanation we ascribe to, suggested by Leeper [1948], is that the information processing decisions made to arrive at the behavior were useful at one time, but now represent an outmoded and nonoptimal solution to the situation at hand.

Section 1.4 Intensionality

One way in which humans act to bring about advantageous situations is to act purposively. Humans use intensional constructs such as beliefs and intentions to order the environment and direct their behavior. If a motivation theory of human behavior is to be complete, it must account for both the direction of behavior by these constructs and the system's actions upon them. The question is: What are the components of an intensional construct and how should the components be modelled? In this section we will explore the psychological definition of intensionality. We hypothesize that an intensional construct has two parts - name and concept.

Logicians and psychologists use the term "intensional" to refer to several phenomena. (We will use the "s" spelling here to distinguish from "intentional", meaning intended. Dennett [1969] uses the same psychological concept but spells it "intentionality".) One use is to define "intensional" as characterizing all sentences that are not termed "extensional" by logicians [Carnap 1947; Quine 1960]. In this use, the "extension" of a description is the class of all things of which the description is true; the "intension" is the meaning of the description. Logic is extensional; extensional sentences follow the rules of truth functional logic.

Intensional sentences have been defined by their logical peculiarities. Chisholm [1957] identified several characteristics of intensional sentences: failure of existential generalization, nonextensional occurrence, no implication of embedded clause or negation, and referential opacity. Chisholm attempted to identify the defining characteristics to develop criteria for psychological phenomena.

The definition we will use refers to the relation of a person to a psychological object (i.e., an object conceivable by the person). This definition has its origins in Frege's [1892] analysis of propositional attitude constructions. Boden [1972] succinctly identifies the concept as follows: "An intensional sentence is a sentence whose meaning involves the notion of the direction of the mind on an object." Further: "An intensional object can be described only by reference to the subject's thoughts, such as his purposes, beliefs, expectations, and desires." For example, a person looking for a pencil is a subject using an intensional construct - a representation of the psychological object pencil - whether he finds one or not. Typical intensional constructs are beliefs and intentions.

Boden [1970] analyzes intensional constructs in terms of internal models. The brain builds representations or internal models of the environment. These representations mediate between stimulus and response to direct the organism's behavior. For example, the concept of hand is an internal model which "resembles" a hand in certain respects but is not a hand itself.

Another example Boden uses is the passerine chick. The passerine chick will crouch the first time it sees a hawk flying. This response requires a discrimination of one particular stimulus class from other. The innate mechanism which is sensitive to this stimulus class serves the function of a model.

Boden does not say whether she considers the chick's behavior intensional. The question is whether stimulus discrimination characterizes intensionality. That is, are the mechanisms that are generating discriminatory behavior sufficient to incorporate beliefs and goals? We argue that they are not. The reason lies in the fact that beliefs and goals represent nonimmediate behavior. We can build a discrimination mechanism using a discrimination net with single-direction pointers. That is, the only access to the stimulus representations is by actually perceiving the situation.

But if the system is to use beliefs and goals, the system needs another access to the internal representations (concepts). For example, the system must be able to manipulate the representation of a pencil so it can establish a goal to find a pencil. We postulate that each intensional construct has two parts: (1) a name and (2) a concept. A concept is sufficient to discriminate a particular situation in perception. A name is an alternate way of accessing the concept. We speculate that names are linguistic in origin. The name and concept representations can access each other, but with different consequences.

Boden alludes to this distinction in her examples of machine behavior. She describes a quality-control machine in a canning factory that distinguishes overweight cans. Boden asserts that this machine can generate its behavior without reference to intensional constructs. (However, the machine still discriminates, so in some sense it has an ability to represent a model of the conditions it detects.) Boden contrasts the quality control machine with a machine searching for a certain library book. She states that this second, goal-seeking machine's behavior cannot be explained without reference to search rules and criteria for recognizing the goal. The explanation includes procedures to search a library and recognize the correct book. These procedures necessarily must be accessible to internal processing by some means other than the desired book's perception if the machine is to direct its behavior towards the goal.

Thus, a discrimination mechanism is not sufficient for all intensional constructs. In this work, an intensional construct will refer to an internal model or representation in either a program or person that has two parts - name and concept. A concept is a stimulus discrimination (perception) or realization of motor responses (action). A name is used to access concept representations without perceiving the corresponding object or performing the corresponding action.

Chapter 2. Overview of the Simulation Model

We shall present a brief overview summarizing the three succeeding chapters.

Section 2.1 Task specification

Section 2.1.1 The interview task

The vehicle chosen for realizing the motivational and intensional constructs is a simulation model of a paranoid person participating in a psychiatric interview. The implementation of the model is part of an attempt [Faught, Colby, & Parkison, 1977] to construct a simulation model of a particular theory of paranoia [Colby 1975]. The model communicates in English with a psychiatrist attempting to diagnose the model-patient. The interview usually lasts about an hour and consists of 30-40 input-output pairs of English utterances. A brief description of the theory-modelling aspects follows.

In general, one constructs a model to make the corresponding theory more acceptable or plausible. Computer simulations, however, have three properties advantageous to the theory-builder: (1) the assertions of the theory must be put into a (relatively) consistent formalism, that of a computer programming language; (2) the assertions of the theory must be made complete, to specify the operation of the program in every possible situation it will be in; and (3) once a reasonable model has been programmed, the simulation can be exercised to discover complex consequences which were not initially noticeable from the assertions of the theory.

The simulation model we constructed attempts to imitate some of the symbol-manipulation aspects of human behavior. We placed special emphasis on representations of purposive actions and affective variables which guide those actions. The advantage of attempting a simulation model at this level is that the theory-builder must specify the details of the theory elements as they interact with the rest of the simulation model. The disadvantage is that the model-builder must include a number of components in the model which have no significance to the theory at hand but nevertheless must be included to obtain realistic performance. In addition, there are a number of requirements and constraints which paranoid phenomena place upon the simulation model.

Our task required a situation in which paranoid behavior could be exhibited. We chose the diagnostic psychiatric interview as the appropriate situation. Participating in a psychiatric interview is an ideal task because it displays so many of the cognitive, affective, and conative processes that the paranoid mode uses in communication. When participating in an interview, a person's preconceived ideas about the purpose of interviews and the typical events occurring in them set up expectancies about appropriate actions. The person's needs and desires, originating from a global self-interest for survival [Faught 1975], motivate his interview participation by specifying goals to accomplish during the interview. The person uses these goals to form plans to satisfy his goals, and he executes actions that make up the plans. He then observes and evaluates the actions that are taking place to determine whether his goals are being achieved or whether he must cope with some new situation. His needs and desires are tied to the success or failure of his actions in modifying the states of both his inner and outer environments. Finally, as his needs change or as his situation changes, he can modify his goals and actions to attend effectively to his needs, thus steering events to his own ends.

The interview situation also constrains the processes in several ways which make the simulation manageable. There are only a few sets of facts about which the model needs to make inferences: the interviewer's actions (limited to linguistic behavior); the model's own immediately

preceding actions; the course of the interview so far; and predictions about the future course of the interview. Further, there are only two participants in the situation, so third-person interpretations and three party interactions are nonexistent.

Not all of the components of the simulation model are crucial to the theory. The simulation model's performance requires that the model participate in an interview using English, that it attempt to satisfy some of its own needs and desires, and that it exhibit paranoid phenomena in its performance. We separate these requirements into three types of simulation tasks: language processing (recognition and generation), purposive behavior, and paranoid processing. While language processing is the least influenced by paranoid phenomena, it is critical in order for a simulation model of human purposive behavior to pass Turing-like indistinguishibility validations [Heiser, et al, 1977]. The elements of purposive behavior (motivation and intensionality) are not always influenced by the paranoid mode, but it is difficult to imagine describing paranoid phenomena without comparing normal states of these elements to paranoid states.

Section 2.1.2 Constraints

Another requirement placed on a model of motivational mechanisms is the set of physical and developmental facts about humans. These fall into three broad categories [Simon 1969]: the survival of the organism in the current environment, the ontogenetic development of the individual, and the evolutionary development of the species. These constraints place limits on the types of information processing possible, and they provide a source of tasks to be considered when extrapolating hypotheses of the original model's potential performance.

The motivation mechanism must continually provide for survival in the contemporary environment. The basic strategies for survival are avoidance of situations or conditions detrimental to the organism's survival, and enhancement of situations or conditions which benefit the organism. If not all of the recognition and response processes are hard-wired into the system from birth, then the organism must learn these processes. The organism associates the appropriate avoidance or approach values with the processes in the form of noxious or satisfying stimuli. Further, the system must have a notion of anticipation of situations to be aware of situations which have a potential for significant events (advantageous or not). The ability to recognize previous situations that were significant enhances this sense of anticipation. The organism must remember the value of these situations so that it can approach or avoid them before the fact. Finally, the physical environment requires a timely response to some stimuli. Learned responses must show the same timeliness as the original programmed responses.

If the organism is to survive changes in the environment, it must be able to adjust the way it fulfills its internal needs from the external environment, assuming it cannot alter its internal needs. When the organism recognizes a source of danger, either by observation or by symbolic communication, it incorporates a recognition of the new situation into its value system so that the system will take appropriate notice when the danger is about to occur. Similarly, the organism develops new actions to deal with the situations. These actions become timely so as to deal with the situation when it occurs in real-time. When compared to other mammals, human ontogenetic development is striking. At birth, humans almost totally lack tropisms (innate or hard-wired cued actions). While mooses can stand within minutes of their birth in the snow, and while baby chicks quickly cower at the form of a hawk overhead, human babies seem to start with a minimal set of tropisms (e.g., rooting and crying behavior). But even at this stage infants have a mechanism anticipating danger. Without this mechanism mothers would never know when their children were hungry or in pain. Finally, before they are able to speak and develop a facility for language, humans are able to cope with the environment by judging stimuli and performing appropriate actions.

Evolutionary constraints provide another source of comparisons. Basically, if we develop a model of human motivation, we should be able to remove parts of it and still have a functional model of motivation that would explain the motivation of other mammals or other animals. By removing human language and naming abilities, we should still have a system that can sense danger, anticipate important events and prepare for them, learn to evaluate new situations and remember them for the future, and incorporate new actions into their repertoire in such a way that they can cope with the new situations in a timely manner.

Section 2.2 Actions

Section 2.2.1 Human actions

If a theory of the motivational mechanisms in human behavior is to be correct and useful, it should account for a variety of human behavior. Let us examine some of this behavior. Humans are purposive beings. They act to bring about situations that they have in mind but do not yet exist in reality. But not all human behavior is purposive. Some actions could be considered mechanical or machine-like, such as shifting gears in a car. A mechanical action serves a purpose but is not itself step-by-step purposive. When one is first taught to drive a car, one is forced to take into account every detail of the actions to be performed. But after a suitable number of task performances, the individual motions no longer consume one's attention. The task becomes one unit of activity [Dennett 1969]. One explanation of the task performance is that one performs the same cognitive processing, only more quickly. The explanation we will elaborate upon is that the entire task has been incorporated, or almost compiled, into a single, unanalyzable unit of activity. The human is not aware of any deeper activity. Any post-performance analysis is an observation of behavior and not a report on internal processes. These units of activity can also be automatic in the sense that environmental conditions dictate their instigation. One could imagine that the person wanted to put the transmission into third gear and shifted the gears with this goal in mind. The explanation we will use is that the conditions in the situation, (e.g., the speed of the vehicle, the whine of the motor) prompted the action of gear shifting. This action may itself not be considered purposive, but it may still be part of a larger purposive act such as going to the market; therefore, it must be explained in terms of the purposive activity of the person. In addition, one must consider the compatibility between the mechanisms that allow the global goal to direct action, and the more local and specific mechanisms that motivate these lesser actions.

Another type of human behavior is similar to interrupts in computers. Suppose a child darts in front of an auto. The driver stops whatever he is doing and forcefully brakes the car. He does not wait for an ongoing action sequence to terminate. He has an interrupt mechanism that allows more important information or actions to take command of the system. There are three reasons why this process is different from a tropism (e.g., fear of falling): (a) The reaction of stepping on the brake to avoid a child is learned at some point after birth. (b) The reaction is situation dependent. A person in a passenger set has not nearly the same tendency to try stepping on the brake as a person in the driver's set. (c) There is an evaluation of the input which occurs, e.g., whether the object was a child or a large doll. One might argue that the reaction becomes as automatic as a tropism. That is entirely possible given that the situation happened often enough to become an automatic response. In any case, the sensory input has the ability to interrupt the action and substitute a different action.

Another action to account for is one that may at first glance be considered a goal-oriented action. One example of this action is a person using the brake to stop a car. The action may be thought of as executing some movement until a condition becomes true and then ceasing that movement. Another example is a person drinking from a water fountain. The person drinks water

until his mouth is full or until he needs air. One could describe the action as establishing a goal to have one's mouth full or a goal to breathe. However, a simpler explanation is that the action of taking in water will be performed until some condition is true, at which time the situation no longer calls for taking in water. In any case, the action needs to be explained, particularly the motivation for ceasing the action.

Other behavior involves the use of symbols within models. As a simple example, take the braking example above, either after the child has run safely across the path of the car or after the doll has been run over. Assuming the auto has not lost too much speed and the gear shifting action was interrupted with the gear shift in neutral, the driver must finish the shifting operation. But the conditions have changed, and the driver must resume the gear-shifting action sequence from a nonstandard initial condition. The driver must perceive the situation, symbolically suggest the appropriate action, e.g., shift from neutral to third gear, and perform the necessary action.

A more typical task using symbols is the formation of a plan to achieve some goal or set of goals. If I am hungry in my office I can proceed to form a plan to get my wallet, walk to the third floor vending machine, and obtain the sought-after gratification. One might say that I performed the plan, but the plan itself was not enough to cause my movements. I can form the plan but choose not to carry it out. Some special mechanism or special use of symbols enables the symbols to cause physical (or mental) actions.

The use of symbols enables behavior that is not possible with only sensory data. For instance, my concept of the machine, not the actual vending machine, caused my actions. If in fact there had been no vending machine, that someone had removed it without my knowledge, my plan would have been the same. Further, I can extend my sensory input using symbols by communicating the symbols to others. I can ask my office mate if he knows whether the machine is still there. Finally, the symbolic information can take on importance to interrupt actions, as in someone telling me about an oil warning light in my car. The motivating mechanisms for these and other symbolic processing should mesh with and take advantage of mechanisms accounting for the simpler actions described above.

Section 2.2.2 Action and Interpretation Patterns

We categorized the behavior of our paranoid model into three major tasks: (1) recognizing and analyzing the situation that the model is in, (2) performing actions that will accomplish the model's intentions, and (3) motivating its own behavior. Recognizing specific linguistic actions and interpreting them based on the linguistic context are necessary in the model's natural language dialog situation. Performing linguistic actions using the appropriate speech act fulfill's the model's intentions. Motivation of the model's behavior stems from a function called affect (emotion). Affect embodies the model's significant experiences (the model-patient's needs, desires, and interests) and continually modulates the recognition of situations and the performance of actions.

The major design decision in the construction of the simulation model was to use a production system (PS) as the basic process. The nature of the task dictated the choice to some extent. One of the characteristics of dialog is that any input or situation may follow any other. Thus, to a large degree, the environment dictates the system's response. Because of this variability in successive situations, the entire set of data to recognize situations and the corresponding potential responses must be available at each step. A PS is a computationally effective method for achieving this availability of information.

The model's external behavior is a set of sequentially-performed actions. Newell and Simon [1972] pointed out that the human information processing system is basically a serial system: it can execute one elementary information process at a time. These actions are modelled in production rules called Action Patterns (APs). The left-hand side of an AP contains the conditions that specify when to perform the action. The right-hand side contains the action to be performed. For example, if the input from the interviewer is "Tell me more", one of the actions that the model can take is to continue on the current story line. This is accomplished with the AP:

(GO-ON → GET-STORY-LINE)

The activated actions (those in the conflict set) are matched against a small set of action rules to group the actions into multiple attributed actions (complex actions which can perform several simple actions). For example, the two actions of answering a question and showing hostility could be combined to show hostility in the answer to the question:

(ANSWER → REPLY)
(ANGER → SHOW-ANGER)
<< SHOW-ANGER & REPLY →→ ANGER-REPLY >>

From the set of multiple attributed actions, conflicts are resolved by choosing the complex action with the highest affect response according to a predetermined ordering (first match). Specific instances of APs dealing with conversation structure are called conversational action patterns (CAPs). Single APs can be linked together to model multi-step sequential actions. For example, a multi-step sequence for responding to a question contains actions to find the answer and then express the answer:

(QUESTION → FIND-ANSWER & T1)
(T1 & ANSWER → REPLY)

(T1 is a tag to link the two patterns.)

The internal behavior of the model, however, falls into two categories: (1) parallel recognition of situations and (2) sequential interpretation of situations and manipulation of beliefs. This distinction is identical to the assimilation/interpretation forms of Newell and Simon [1972]. Interpretation Patterns (IPs) accomplish parallel recognition (matching) of situations. Action Patterns (APs) accomplish sequential interpretation of situations and manipulation of beliefs. The output of IPs are further conditions which serve as inputs to APs and other IPs. For example, if the input from the interviewer is a statement and fear increases at the same time, then the input can be interpreted as a threat:

(STMT & FEAR-CHANGE : THREAT)

The production system execution is based on a five-step process:

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

(2) Activate - All of the APs and IPs (PS rules) whose initial conditions (left-hand sides) are true are activated. All of the right-hand sides of activated IPs are added to the set of true initial conditions; these new conditions may activate additional rules.

(3) Match - All of the action elements of the activated APs are matched with a set of action rules to combine the action elements into as few multiple attribute actions as possible. One action is then selected (conflict resolution) to be performed.

(4) Execute - The selected action is executed.

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

Section 2.3 Motivation

Section 2.3.1 Human affects

In the previous section we examined some possible types of human actions. In this section we shall see how affects (emotions) influence those actions.

There is a small number of fundamental (primary) affects in humans - probably less than ten [Tomkins 1962]. We can identify these affects by their typical facial and bodily response in humans [Izard 1971]. The function of affects is to motivate human behavior. Each affect can be activated individually and can vary over a range of activation levels. High activation is indicated by excessive facial and bodily response and motivational influence; low activation is indicated by little response and influence.

Affects influence actions in several ways. Affects provide input stimuli to other processes by adding information about the stimuli's significance to the person [Solley & Murphy, 1960]. Affect arousal arises from two sources: sensitization (stimuli from the outside), and anticipation (association with previous experiences in similar situations). For example, a particular environmental situation may activate fear, e.g., giving a speech. Concluding new beliefs may also activate affects, e.g., concluding a best friend is a narcotics agent. Affect activators themselves have a dual function: (1) to arouse the person's responses and (2) to cue the correct response or action. Once an affect is activated, it may trigger a response typical to that affect. However, its response may be tempered by the current situation. If several appropriate responses to the environment exist, the affects can establish priorities among the responses by choosing one to be performed first. However, one affect can monopolize the person's ongoing action [Averill, Opton, & Lazarus, 1969]. Also, an affect may interrupt the ongoing action [Simon 1967]. Finally, affects enable humans to be self-initiating: Humans need not wait for a stimulus (from the environment or a drive) before acting.

Let us examine the specific situations or factors that affects must measure.

For a person to survive, he must have a certain minimal set of motivational constructs guiding his behavior. Foremost in this set is an appraisal of danger or objects and situations to be avoided. The appraisal of danger involves a set of stimuli to be avoided and a global processing design which causes the person to value most highly the lack or removal of such stimuli. In humans, this set of stimuli comprises the mechanism of pain. The state of a person that has the highest value with respect to the criteria for decision making is the state in which painful stimuli are nonexistent or at least diminishing. In addition to suffering immediate pain, humans can anticipate pain. They appraise themselves to be in danger of suffering from further undesirable stimuli. This anticipation is the concept of fear. Distress is the affect triggering mechanism that brings danger to the person's attention. Distress must change the person's state radically enough for the person to take note of the danger. In addition, if the person fails or is thwarted in some act, he must appraise what seemed to be the cause of that failure so that he can bring his resources to bear upon the prevention of future failure. This failure or thwarting of anticipated gain, or unanticipated failure, is the basis for anger in humans.

A person is able to fend off, with fear, anger, and distress, harmful or at least undesirable stimuli. But the person needs more affects if he is to act on his own to procure beneficial situations. When the person attempts to cause or increase certain stimuli, those stimuli are valued as desirable. Of course, if a person responds to the same stimulus by both avoidance and approach, the competition may thwart both actions.

These measures of stimuli, then, are the bases of the motivational constructs for self-sufficiency and self-enhancement. However, these measures or criteria and their associated approach or avoidance values are not enough. The person must apply the criteria to specific stimuli to direct the person's mental processing. The process of associating the criteria with stimuli becomes a simple form of learning. The information processing routines then must promote behavior which, with respect to the associated criteria, will enhance desirable stimuli and diminish undesirable ones.

However, in order for the motivational mechanisms to be most effective at guiding behavior, the mechanisms should influence decisions at the most primitive level. That is, not only should affects be able to influence the selection of an action response, but they should also influence cognitive recognition of situations and inferential processing.

Section 2.3.2 Affects

The model's affect mechanism contains six discrete affects, each of which can vary in activation level (strength). The six affects are characterized as either positive (joy and interest), indicating conditions favorable to the quality of the organism's survival, or negative (fear, anger, shame, and distress), indicating detrimental conditions. Each affect measures a distinct condition. For example, fear measures anticipated pain; anger measures the failure or thwarting of anticipated pleasure. The affects are represented by variables which are separate from (but available to) the PS.

Affects have activators and responses. The activators are those conditions which raise the affect's level. These conditions cause the physical and phenomenological reaction characteristic of the particular affect. The responses are mental processes which the affect's new higher level activates. Many of the activation and response mechanisms stem from the fact that affects may be named in the elements of APs and IPs. If an affect name is an element of a right-hand side of an AP, the AP is one of the secondary activators of the affect. Similarly, if the affect name is an element in the left-hand side of an AP, the AP is one of the possible responses to the affect. Of course, the other elements of the patterns mediate the processing that the system takes.

Primary activation of affects is due to an interpreter mechanism which models the anticipation of future situations using the APs. For example, anticipated pain is the primary activator of fear. Whenever an AP having a pain element in its right-hand side is activated, the fear affect also is activated.

The affect system also has an interrupt capability which stems from the constant activation and deactivation of patterns. For example, while a set of APs with low affect conditions named in their left-hand side elements is being processed, the system may discover some new fact which raises fear. The interpretation of the situation (residing in IPs) changes immediately, and a new interpretation, that the situation is threatening, appears. The old interpretation has been deactivated since the left-hand side conditions are no longer true. The system has effectively been interrupted from its previous processing, and processing based on the new interpretation begins.

Of course the affects can influence any processing, since affects may be on either side of the

patterns. Affects may select inference rules or beliefs to explore, select goals or intentions, or actions to satisfy goals. In particular, they may alter the interpretation of situations by residing in the left-hand sides of interpretation patterns.

Section 2.4 Intensionality

Section 2.4.1 Human uses of intensional behavior

Another distinct characteristic of human behavior is the ability to construct and use internal models of the environment. Such behavior is termed intensional. A human behaves intensionally by constructing internal symbolic representations, or models using these representations, to guide his actions. The representations may be used either to interpret external phenomena, including the person's own behavior or to prescribe actions for the person to perform. That is, internal models may be used either to impose order on the vast amount of sensory input and cognitive representations, or they may be used to impose order on the large number of sequences of actions which may be activated to deal with a situation.

The most commonly thought of use of internal models is to simplify an otherwise complex representation. For example, optical sensory input arriving at the brain is a representation of what the eye sees. To translate this input into a different representation is not to metaphysically alter the sensory input's representational quality. The first metaphysical alteration is at the point where the original light ray strikes the perceived object in question and is modulated by the surface of the object. At that point, the light ray, or rather the package of light rays, contain a representation of the object which reflected them. It is the task of a sensory input device, such as the eye, to transform that representation into electro-chemical impulses usable by the brain. One use of internal models is to limit the sensory data into the brain by forcing the data to fit into an internal model which accounts for the data. The brain then uses the internal model to restrict further input by forcing the input to fit the given model. The input data may originally suggest the internal model, but the brain locates the internal model. The model tends to replace and thus simplify the original data in further processing.

Data simplification can involve objects, object properties, or scenes of objects. One can perceive a wall's material, its redness, or its relation to the other walls in a room. In perception, simplification arises from the ability to ignore background information, knowing that it is accounted for. There is a further simplification possible involving the abstraction of known objects. When a person perceives an object, he does not have to compare this particular object with all other similar objects he has known, no matter how represented. Instead, he matches the object to a stereotypical object and makes generalizations from it. Thus, internal models make possible the information reduction found in stereotypes.

One further simplification is possible, relating to the metaphysical oneness of spatial objects. If I leave a package of gum in my desk overnight, I believe that the package I pick up from the exact spot the next day is the same package. If someone has substituted a different package and tells me so, then metaphysically the new package is different from the old, no matter how similar in appearance they may be. Internal models make the representation of this oneness possible, no matter how different the sensory input signifying the same object (as from different angles) may be.

Note that these simplifications may also exist in nonhuman organisms, and therefore are not dependent upon language as we know it. Animals are able to simplify visual scenes and restrict sensory input, especially from locations that are familiar. Similarly, animals are able to generalize stereotypical objects, e.g., the passerine chick. Finally, for a dog to know its master, it either has (1) a

representation of unique objects, or (2) a complete enough description of one certain object to distinguish that object. These simplifications seem to be possible by organisms without language aptitude.

The question then is what enables humans to manipulate internal models or to build models of models? The other use of an internal model is to represent a concept by name. We shall use the term "symbol" to refer the name of an internal model. A symbol can be a handle or name for the phenomenon it represents. Suppose I have seen many brick walls in my life and whenever I perceive one I model it with an internal model. We will call this model "G0123" and define it as being only accessible by seeing a brick wall. (Note that "G0123" is OUR label for the internal model and not the system's label.) Suppose I am now looking at a brick wall. My visual field representation is greatly simplified by the data from the brick wall being associated with an internal model G0123. However, if the model G0123 has no label attached to it (name which is accessible to me), than I have no way of referring to the concept of brick wall.

If, instead, I associate the symbol "BRICK WALL" with the model for brick wall, then the symbol "BRICK WALL" itself becomes a sensory input, subject to internal models and associations with other symbols. This is the crucial step. A symbol or name of a concept must have a physical reality of its own. Symbols are included with other sensory input. The normal perception and model-producing mechanisms can be applied to them. In humans, this takes the form of verbal language communication.

The tie from symbol to internal model must be traversable in both directions. Triggered by sensory data, the internal model suggests the symbol to represent it whether the sensory data is from external sources or internal sources representing an action. In the opposite direction, a symbol needs to access the internal model which it names. For example, if I am looking at a picture that I cannot comprehend and someone suggests that it is a brick wall, then I must access the representation of the stereotypical brick wall and apply it to the picture. Similarly, if I want a drink of water because I know I will have to do without for a few hours, I must symbolically direct my behavior. This direction occurs even though there may be none of the usual stimuli which usually trigger my water obtaining behavior. With this tie from symbol to internal model, humans can communicate information about the world and learn facts without ever having to experience them.

We examined an internal model without a name in Section 1.4 and discovered that the model could be used to discriminate but not to represent nonimmediate data. Now we have the question of whether symbols (names) can be used without some attachment to an internal model (concept). To say that a symbol has no internal model which it names is to consign the symbol to place-holder status. In this case, the symbol may have properties attached to it, but it remains only an anchor for these other links. Suppose the symbol has a corresponding internal model, but the model comprises only symbols and not sensory input (beyond the sensory properties of the name itself). The bottom level symbols will be devoid of representation and therefore lacking in attachment to any external mechanism. It is difficult to see how such symbols could be used by a system to its advantage.

Symbolic models are useful to human behavior due to the models' ability to simplify, make representations of reality, and provide handles for symbolic communication. An interesting problem is determining the motivational mechanism for encouraging humans to use symbols effectively. If we assume that person has symbolic representations of the external world and actual sensory input from the external world, then we must assume that the symbolic model does not always accurately reflect the world. Similarly, if we assume that the person has actions that he can perform on the world and representations of those actions, we must assume that the models of actions are not always accurate. Sometimes the person is not able to perform in accordance with his model, or the projected outcome does not happen. Thus, some measures of how well these internal models match the actual reality of

the individual's relation to the world would reflect the person's ability to construct and use these models. In some sense, these measures reflect the person's ability to control his external environment. If the person can be convinced that controlling and changing his outside environment is to his advantage then a motivational mechanism for encouraging the use of symbols can emerge.

As to the necessity of intensional behavior in an intelligent system, clearly systems can survive and enhance themselves without symbolic models. However, if we wish to consider an system that has the ability to listen to instructions or descriptions of the world symbolically, then symbolic models of the system's representations are necessary for the system to perform.

Section 2.4.2 Intensional constructs

In addition to the AP-IP system, there are two examples of intensional constructs in the model: beliefs and intentions.

The intensional constructs are realized in two constructs: a concrete conceptual form for recognition of the concept based on experience with it, and a symbolic or token form for labelling or naming the concept. A token is a linguistic name. (We use the word "token" similarly to the computational linguists' type-token distinction. A concept is analogous to a "type".) Token forms can be reported and manipulated directly by the system; concept forms cannot. Concept representations are APs and IPs. For example the concept representation of an insult is the IP that interprets an input statement such as "You creep" as an insult. A symbolic representation of an insult is necessary for the transformation of a statement such as "He insulted you" into the appropriate anger. Internally, a symbolic representation is partially linguistic in nature. It is an IP with the linguistic representation of the symbol as the IP's left-hand side, and the node in memory corresponding to the symbol as the IP's right-hand side.

The processing done on the two types of representations is what gives one type the opportunity for representing nonimmediate data and the other not. Concepts are subject to the rapid activation and deactivation of all APs. Concepts are active whenever their initial conditions (from direct immediate input) are satisfied, but are not active when their conditions change. Thus, concepts are only useful for processing immediate reality. Tokens, however, are rehearsed in short-term memory. They can be active without their associated initial conditions being true. Tokens serve as intensional place-holders in representations because they can be activated based on their linguistic name, independent of the sequence of concept conditions that called them. This gives tokens their power to act as cues for missing or nonimmediate data.

Beliefs and intentions, the two intensional constructs in our model, guide the system's behavior by participating as elements of APs and IPs. Inference rules are implemented in IPs. They have beliefs and intentions as potential consequents that can be set true when all of the left-hand antecedent elements of the IP are true. Beliefs can also be left-hand elements of IPs for establishing a context to interpret a situation. Intentions can be left-hand elements of APs to serve as intensional actions or to establish subgoals. As always, affects can be elements of either side of the patterns and thereby participate as antecedents to inference rules.

The model uses intensional constructs when concrete situation-action rules fail. Although actions which originally depended upon symbolic intensional models can become concretized, the only truly intensional behavior occurs when an internal model is accessed by its name. Intensional constructs are especially useful for analysis of action failure, perception extensions, and symbolic (as opposed to experiential) learning.

Section 2.5 Implementation

This section describes the model's programming and system considerations. The model was implemented in the programming language MLISP, a dialect of UCI-LISP, at the SUMEX project at Stanford University under a TENEX timesharing system on a PDP-10 computer.

The program described in this work is one of three programs in PARRY3. The other two programs are the language recognizer and the language output program. The language recognizer and the model each require approximately 200 pages (512 36-bit words per page) of core. A typical response requires 1-2 seconds of computer time for the recognizer and 1-3 seconds for the model. The language recognizer and the model run in separate forks so that the model may continue processing while the recognizer is accepting and translating the next input. Of course, the model is free to generate further output and in fact on occasion interrupts the interviewer's typing.

The language recognizer [Parkison, Colby, & Faught, 1977] provides the language interface from interviewer to model. It translates the English input utterance into a semantic representation of nested predicates and a set of variables. For example, the following input utterances on the left are translated into the nested predicates and variables on the right:

```
"WHY ARE YOU IN THE HOSPITAL?"
  → (REASON (LOCATION I HOSPITAL) ?) QUESTYPE=WHQUES
"ARE YOU IN THE HOSPITAL?"
  → (LOCATION I HOSPITAL) QUESTYPE = YESNO
"YOU ARE IN THE HOSPITAL."
  → (LOCATION I HOSPITAL) QUESTYPE = STMT
"WHERE ARE YOU?"
  → (LOCATION I ?)  QUESTYPE = WHQUES
```

The semantic structure is matched to the memory of stored conceptualizations (approximately 650 ideas that the model can process). Some conceptualizations have slots that can match any element or have elements which can be generalized. For instance,

```
(EMOTE I A)
```

matches any of the following structures:

```
(LIKE I MAFIA)
(ANGRYAT I MOTHER)
(FEAR I MAFIA)
```

The model can partially recognize a large class of inputs using these general conceptualizations. If the model has some information on the subject, it gives a related answer. In the EMOTE example, the system uses a special AP for answering this type of question to access the matched concepts (e.g., MAFIA) and to detect whatever affects changed. Then the system reports these affects as the model's reactions to the concept in question.

The language recognizer provides a number of other facts about the input. The recognizer supplies a list of embedded clauses and nouns so that the model may react to the concepts with its affects. For example, the input

```
"DO THE NURSES BELIEVE YOU ARE CRAZY?"
```

translates into

(BELIEVE NURSES (BE I CRAZY))

with the embedded clause

(BE I CRAZY)

and the embedded nouns I and NURSES. These concepts activate affects in the model. For instance, reference to the self being crazy (BE I CRAZY) raises shame.

The recognizer also provides information on the interviewer's style. For example, the recognizer calculates the level of confidence in the recognizer's translation based on the number of symbols in the input expression that could not be recognized as words and the number of spelling mistakes that could be corrected. The recognizer determines the input's amicability according to the presence or absence of positive concepts ("self is correct") and negative concepts ("self is dumb"). The recognizer also provides details of the subject and auxiliary form of the input utterance so that a generated elliptical output can match the input.

The language output program is small and simple. It contains a set of procedures for producing a formatted natural language output from a semantic representation of a concept plus some flags.

One of the problems in computer simulations of complex mental functions involves making it possible for the simulation to show its internal subtleties through its external behavior. In our simulation, linguistic output expressions should perform several functions at once. For example, the answer to "WHY ARE YOU IN THE HOSPITAL?" may contain an informational belief "BECAUSE I AM UPSET", and at the same time carry affect content relating to the interaction of the conversational participants, e.g., "WELL, I REALLY THINK IT'S BECAUSE I AM UPSET" or "YOU SHOULD BE ABLE TO TELL I AM UPSET". The solution in the current model is to have each linguistic output perform only one function. Each linguistic output must either convey specific information or convey affect. Each response, however, may contain more than one output expression. The output expressions are chosen from a stored list of utterances which all express the same information or affect.

The language output is embellished slightly, however. For YESNO questions such as "DO YOU TAKE DRUGS", an elliptical answer such as "NO, I DON'T" is generated in addition to the normal output "I DON'T TAKE DRUGS". Depending upon the type of affect state that is being exhibited, the two answers may be concatenated or one answer may be chosen over the other.

The output program works with an online data file of about 50 pages containing approximately 1800 output utterances.

The model's basic construction is a production system with approximately 300 rules. (Appendix B contains a list of the model's rules.) Of these rules, 100 are APs and 200 are IPs. The PS interpreter requires about one half second of computer time per interpreter cycle. Two to five cycles occur between each input/output pair.

The general form of a rule is:

(Pnnnn NAME PAST : PRESENT → ACTION FUTURE)

Pnnnn is an internally generated LISP name. NAME is the logical name for the action when used as a subaction. PAST is a list of left-hand side elements tested by Activate. PRESENT is a list of

conditions to be set true by Activate. ACTION is an action that the Match procedure tests when forming multiple attribute actions. FUTURE is a list of conditions that the Bookkeep procedure sets to true if the ACTION is successful.

Of the 900 rules, approximately 70 are generated from multiple step rules (MPATs). An MPAT rule is a rule that spans several APs. For example, an MPAT rule for answering a question is:

```
(M0810 :(QUES) !1((FINDANS INPUTF))
       FINDANS SAYANS !2((OUTPUTF SAYANSDN)) )
```

QUES is the initial condition to the entire pattern. QUES is true if the input is a question. The two actions of the MPAT are FINDANS and SAYANS. The ((FINDANS INPUTF)) expression specifies that the input to the FINDANS action is INPUTF. The ((OUTPUTF SAYANSDN)) expression sets the results of the action SAYANS to the variable OUTPUTF.

The MPAT above would be translated into the following single step rules:

```
(P0811 FINDANS ((N0811 NIL) ITALK QUES) :
  ((FINDANS INPUTF) FINDANS) → (FINDANS) (N0811))
(P0812 FINDANS (N0811 ITALK FINDANSCN (FINDANSDN NIL)) :
  (FINDANS) → (FINDANS) (N0811))
(P0813 SAYANS (FINDANSDN N0811 ITALK) :
  ((SAYANS FINDANSDN)) →
    (SAYANS) ((OUTPUTF SAYANSDN) N0811))
(P0814 SAYANS (N0811 ITALK SAYANSCN (SAYANSDN NIL)) :
  ((SAYANS SAYANSCN)) →
    (SAYANS) ((OUTPUTF SAYANSDN) N0811))
```

There is a separate program which reads in these MPATs and breaks them into a series of single-step rules, linking them with generated tags. There are approximately 50 MPATs.

The model's set of beliefs and inferences are a subset of its conceptualizations. The model could use any conceptualization as a belief. In practice, however, the model uses only about 100 conceptualizations as beliefs to represent facts about the interviewer and the interaction (Appendix D). In addition, the model uses approximately 120 patterns as rules of inference to manipulate beliefs (Appendix E). The model has about 30 distinct actions that it can perform (Appendix C).

The performance demands upon our model are that it take no longer than 5 seconds of computer time, 15 seconds of real time to respond, and that it respond to an unrestricted set of English input. The program required some optimization in its pattern activation. For example, the innermost loops of the activating mechanism were machine language coded. Also, to reduce page faulting, most of the left-hand elements of productions rules were placed in as few core pages as possible.

Chapter 3. Review of related work

In this chapter we will discuss other related research. This research falls into three broad categories: problem solving systems, systems that explain actions and plans, and emotion-motivation models.

Section 3.1 Problem solving

The first bench mark of AI work in problem solving was the General Problem Solver (GPS) developed by Newell and Simon [Newell & Simon, 1963]. GPS was a general paradigm for human problem solving behavior. The research goals were to both perform problem solving behavior adequately and simulate the way humans perform such behavior. GPS was applied to problem solving tasks having objects (states in the world) which could be transformed by operators. The major specification of the task was the goal to be achieved and the task's component subgoals.

The top level execution of the program was to compare the end object or top level goal with the current object and try to reduce the difference. The three types of rules to accomplish this reduction were: (1) transform object A to object B, (2) reduce difference D between objects A and B by selecting an operator, and (3) apply an operator to object A. Each rule had initial conditions to test its applicability. The heuristic for selecting subgoals was to order the list of subgoals with respect to differences, and then try the most difficult (largest difference) first and progress to the easier ones. This difference information was stored in a table of operators and differences. The table constituted a taxonomy of the domain. Hence, the system was at its best in domains with well-structured taxonomies.

GPS had the attractive feature of globally deciding what to do next. It appeared to be a candidate for a general motivation system. However, being a goal-oriented paradigm, it could not account for environment-dictated or affect-dictated actions, or for interrupts from competing goals. Also, in order to have one top level goal, a model would have to establish a goal of "exist" or "satisfy self" with operators to reduce the difference between the current state and this vaguely specified state. Besides the unappealing nature of such a goal, in practice it would be inefficient. A model's environment is always changing; therefore a model must continually alter the specification of the top level goal that is possible in the current situation. In reality, the environment is dictating the action. Also, it is computationally more efficient to dictate actions by environmental conditions rather than comparison of differences if the action is always to be the same in a certain situation. That is, sometimes people don't know the outcome or desired state resulting from a particular action; they just know it is appropriate, e.g., saying "hello".

STRIPS [Fikes, Hart, & Nilsson, 1972], was also oriented towards problem solving. Like GPS, its processing was directed by a single goal and used a means-ends analysis strategy to generate a sequence of operations that transformed the initial state to the goal state. Recent extensions [Sacerdoti 1975a, 1975b] have also focussed on the same task of plan formation, given a top level goal to be achieved.

The recent monumental work by Newell and Simon [1972] was also concerned with problem solving behavior. It had a more general goal of describing a theory of the human information processing system (IPS) as applied to problem solving tasks. Their conclusions about the important characteristics of the IPS were:

(a) It is a serial system.

(b) It has a short term memory (STM), a long term memory (LTM), and an external memory (EM).

(c) The capacity and speed of the memories can be estimated.

(d) The processor is a production system, whose conditions for evoking a rule are the appropriate symbols in STM and EM.

(e) The IPS uses a set of symbolic goal structures to organize problem solving.

The important components of their work for our work were:

(1) the plausibility of a production system as the single mode of processing in their model;

(2) the serial quality of symbolic processing, while retaining the parallel nature of memory recall and activation of rules; and

(3) the distinction between assimilation and interpretation in symbol manipulation by the system.

Assimilation refers to the incorporation of symbolic operations into executable production rules. An operation has been assimilated if it is available as a (sub-) production system or set of rules. Interpretation refers to an alternate mode of operation: the controlling PS is an interpreter and the symbolic operator is a symbolic structure that must be interpreted to be applied to the task at hand. The tradeoff between the two is one of efficiency. Assimilated rules need no stacks or external memory to record working through a rule. Interpretation requires working through the rule symbol by symbol, using both time and space resources. However, the symbo. interpreter is general enough to handle any rules of the same form. As examples the authors used rules for manipulating logical expressions. In our work, the assimilation/interpretation distinction will be seen as similar to the distinction between nonintensional and intensional processing.

Section 3.2 Recognizing plans and actions

Plan generation is a more proper subset of our problem than plan recognition. The motivation of an autonomous entity deals mostly with first person recognition of situations and performance of actions and second person recognition of situations. Research on plan recognition, however, deals mainly with third person interpretation of actions in which the model is not a participant in the situation. Nevertheless, there are a few systems which are interesting because of the attempts at developing taxonomies of common sense situations in which people interact and communicate.

Schmidt and D'Addamio [1973] attempted to construct a representation of a taxonomy of third person intention and action. The taxonomy was distinctive in its attempts to pay special attention to the interpreted motivation (e.g., hedonism) of another person's actions. Schmidt and Sridharan [1977; Sridharan & Schmidt, 1977] introduced a plan schema to represent the observed actions to be interpreted as being part of a hypothesized plan. They noted the importance of expectations in interpreting actions as being part of failed plans. Developing an interpretation in their system consisted of constructing an Expectation Schema (ES), matching individual actions into the ES, and computing violations of constraints on the interpretation for best fit. One potential application of Schmidt and Sridharan's work is to simulations that modify their behavior by modeling and interpreting their own actions.

Rieger's work [1976] was another example of taxonomy development. He attempted to find a minimal set of primitive cognitive links that one could use to build large structures of events, state, and state changes. The structures would then be an adequate formalism for expressing human plans and actions.

Schank and Abelson [1975] have attempted to construct taxonomies of stylized everyday situations using their notion of script. A script describes an appropriate sequence of events in a certain situation or context. One of their systems accepted natural language input. The system generated possible explanations of the situation by proposing scripts to be matched and by filling in slots in the scripts. Like Schmidt and Sridharan, Schank and Abelson's domain was the third person interpretation of situations. Charniak [1975] and Bruce [1975a] also fall within this framework. Script-like structures are necessary in our work and are used in a special way. Instead of mapping them on to input data describing a situation, the model must perform some of the actions as a participant in the scripted scene.

Discourse understanding has been one taxonomic task area. Bruce [1975c] recognized that the purpose behind an utterance is necessary for complete speech understanding. He proposed a number of modes of interaction built up from intentions (e.g., edit, resolve conflicts, query, and confirm), all as part of a data base management system. He then organized the elements of the modes into a user model and a task model. Similar modes of interaction are implemented in our work as action patterns. Winograd [1977] presents a "schema" (frame-like) representation for understanding natural language discourse. He analyzes the various discourse components and shows how they could be simplified in a schema representation. Deutsch [1974] focusses on the structure of task-oriented dialogs.

Scragg [1975] attempted to construct a representation of actions which would allow performance of the plans and actions as well as reasoning about them. He suggested a net representation called a "knowledge of procedures" net. The basic unit was an event, a node to which all information about the event is attached. Each event had a START node which pointed into a partially ordered graph of substeps of how to accomplish the event. A substep was either a goal or an action specification. Reason and method links were also available for answering questions about the events. The network was based on a semantic net representation (SOL) by Norman and Rumelhart [1975].

There were two control processes in Scragg's system to direct action performance. The task driven process insured that the goal of the task was accomplished. The motive driven process set up constraints or criteria for how the task got done, e.g., quickly. An interpreter performed the actions by starting at the START node and following nodes. The interpreter put elements on a list for subtasks or wait tasks (those which required the outer environment to respond), according to the partial ordering of the net.

Scragg's model was an attempt to account for different kinds of actions and provide a framework for associating knowledge about actions with an executable representation of actions. The strength of the model was in having parallel specifications for two modes of action - goal (invoked by an explicit goal) and rote (invoked by a situation). Difficulties in the paradigm stem from the lack of generality in the system. Much of the information seemed to be represented in an ad hoc fashion to be used for a single purpose. Also the model performed in the usual slave mode of a question-answering program.

Another related problem is that of generating an action given the conflicting desires, intentions, and knowledge about appropriate actions of a system. Bruce [1975b] addressed this problem with his Social Action Paradigms (SAPs). He used SAPs to explain natural language generation as an action in a social context. In particular, he pointed out that intentional actions can be encoded in presuppositions, linguistic conventions, and discourse structure.

Clippinger [1974, 1975] developed a working computer model of discourse generation (ERMA). He described natural language output not as a process of generation but as a regulation of competing behavior generators. The five major processes in his system each had their own programs and data and their own criteria for what should be said. The processes could inhibit each other or interrupt each other in the middle of output generation. The resulting output mimicked human output behavior in which several competing sides of the person are in conflict. This type of model should be useful in future simulations of human psychopathological behavior.

In an attempt to capture the theoretical organization of actions into coherent plans, Miller, Galanter, and Pribram [1960] proposed a test-operate-test-exit (TOTE) construction as the fundamental unit of human behavior. The TOTE unit carried the assumption that all actions locally regulate their own behavior and insure their own fulfillment or stopping conditions. In this paradigm all plans were built up from TOTE units. TOTE units were hierarchical to allow for subgoals and subplans. The problem with the paradigm is that while some actions are best described by a loop execution or condition terminated action, not all actions fit. Also, the information for action selection in their paradigm was not located with the action but in another set of structures called images. These Image structures were not associated with the TOTE units. Finally, there was no mention of how the plans were motivated.

Section 3.3 Simulations of motivation and belief systems

There have been a number of attempts to simulate emotions or motivational qualities of human behavior. Most of these systems have had some explicit representation of the motivation mechanism and have attempted to make the system autonomous - i.e., directed by some internal measures of performance rather than by an externally supplied goal or command.

ALDOUS, a model by Loehlin [1968], was an attempt to construct an autonomous model of approach-avoidance behavior. The model accepted inputs representing objects and reacted to them by approaching or avoiding them. It had a basic loop of three steps: recognize the object or situation, react emotionally, and act. The model had three affects: love, anger, and fear. The input objects were the numbers 0-999. The model determined the long term consequences of interaction by examining the current emotions. It then modified the long term "attitude" towards the object, i.e., the emotional attribute of the object.

Loehlin's model can be thought of as a single attribution function, although based on its attribution it could approach or avoid objects. Its strength was that it remembered attributes based on both current emotions and past associations with the object. Using this association mechanism, it could model simple conditioned learning. However, all actions were single events; there were no patterns of behavior or intentions. Also, there was no interpretation of events; each event (object) was totally unambiguous. Finally, the model had no mechanism to avoid impending situations that would cause negative emotions or approach positive situations. The model's approach and avoidance behavior was totally embodied in after-the-fact reactions to the objects. Thus, the motivational constructs in the system were post-event determinants of behavior and didn't carry the anticipatory power of emotions.

Doran [1968; 1969] developed a model of a "pleasure-seeking" automaton that executed actions and observed their outcomes. The model selected a goal, planned a series of actions, and performed them. It then observed the environment and compared the predicted outcome to the actual outcome. If its prediction was correct, it did nothing to change the current action plan. If its prediction was not correct, it replanned and generalized the unexpected result. The model had a single positive affect to maximize but no negative affects to avoid. For each variable (representation) of a

state of the environment), the model kept a goal state of how to move to a more positive state. The model performed its execution until satisfied (the affect greater than some threshhold), at which time it suspended its own operation. The model was basically an S-R paradigm, but with the interesting feature that it observed its own actions and corrected both its short term and long term behavior. Once it had observed an action, the model could apply previous actions in equivalent states (according to a generalization scheme) to predict a better action for the current situation.

Kilmer, McCulloch, and Blum [1969] proposed a model of the reticular formation of the brain. (The reticular formation is located at the base of the brain in humans and concerned with arousal and attention.) They proposed that the formation's primary function is to switch from one mode of behavior to another. They postulated a number of distinct behavior modes (e.g., sleeping, eating). The model had a set of independent modules vying for attention until a majority decision was reached as to mode. Positive and negative reinforcement were also included. They could model simple generalization, classical conditioning, and trial-and-error learning behavior. As a theory of affect, their model did not allow for cognitive concepts; the model could only alter the mode of the system.

Kiss [1975] outlined a more general model of motivation. His model characterized motivational processes as assigning priorities to competing goals, and allocating resources of computational work. The report had no detailed proposal for a mechanism, but was mostly an outline of the factors to be considered.

Shaw [1975] outlined a method of dealing with motivational issues in frame-style systems. His proposed model activated frame structures from both high level goals and low level environmental events. The two activation sources interacted in recognition tasks to iterate between hypothesis-formation and confirmation.

Previous work on belief systems by Abelson [1963] and Colby [1964] established the principles of: (a) associating affects with beliefs, and (b) a model responding to its own internal distress or conflict due to its beliefs. In Colby's [1964] model of neurotic processes, a pool of conceptually related beliefs was searched for conflict-generating beliefs. Then strategies were applied to reduce conflict. The transformations attempted were similar to those categorized by Suppes and Warren [1975].

Colby's later work on simulations of paranoid processes (a direct ancestor of this work) was concerned with two tasks: (1) incorporating a specific theory of paranoia into a simulation model, and (2) developing an adequate set of normal and peripheral processes (language recognition and generation) to serve as a base for the theory.

The original model, PARRY1, [Colby, Weber, & Hilf, 1971] had an S-R paradigm as its base. The stimulus was the interviewer's linguistic input from which keywords were extracted to determine the response. A sequence of response processes scanned the input to locate information to which the processes could respond. The sequence was: delusional responses, self-referent topics, flare topics, topic elaboration, and question-answering. Response processes could generate a linguistic output and set affects to new levels. The strength of the model was that the paranoid mode could be so directly expressed and mimicked. However, this property was also a weakness. The model had little normal mode processing. Also, the model's only change in state during the interview was the change in levels of the three affects and lists of topics mentioned in the interview. In spite of these drawbacks, the model was able to pass a Turing-like indistinguishability test [Colby, et al, 1972].

The second model, PARRY2, [Faught, Colby, & Parkison, 1977] tried to eliminate these weaknesses in a number of ways. First, a natural language recognizer was developed [Colby, Parkison, & Faught, 1974] which translated the English input into one of about 500 stored conceptualizations.

Second, a mechanism for building a model of the interviewer during the course of one interview was implemented in a belief system with inference rules. Third, the paranoid processing was separated (somewhat) from normal processing. The number of normal mode actions increased by adding new intentions and actions. The drawback to this model was the multiple ways in which the various mechanisms were implemented. There was no simply-designed mechanism that performed all the processes; each process (inference, affect, and intention) was distinct and relatively uninvolved with the others. Also, the top level loop of the program was an Input-Compute-Output loop. We may fit any theory in such a loop, no matter how badly we have to contort the theory. However, it is more plausible to have the top level motivation of the system stem from motivational constructs in the model, because then the theory is playing a central role in the operations of the model. This model passed a more rigorous Turing test in which the interviewers were asked to distinguish a human from the model [Heiser, et al, 1977].

Chapter 4. Action and Interpretation Patterns

In order to fulfill desires and needs, human beings perform actions which tend to occur in patterns. Actions occur in patterns so that humans may cope efficiently and effectively with the wealth of information that surrounds them. One common example of everyday action performance is the ringing of a telephone. The action sequence is to walk to the telephone, reach for the receiver, pick it up, and say "hello" into the mouthpiece. This is an example of a patterned action - a sequence of individual actions that tend to be performed in the same order, and that once started tend to run to completion of the sequence (unless interrupted by some higher priority process). An example from conversational interaction is the sequence:

(Question → Clarification question → Clarification → Answer)

as in:

"How do you like your work?" - Question

"Why do you want to know?" - Clarification Question

"I was concerned that you were upset." - Clarification

"Well, it's not too interesting; I look forward to leaving at night." - Answer

Other examples of patterns actions in dialog are arching and chaining [Mischler 1975]. Arching occurs when a response to a question is another question which leads the first participant to the answer. For example:

"Do you have a job?" - Question

"Don't most people have jobs?" - Question

"Yes." - Answer

Chaining occurs when the first person asks a question, the second person answers, and the first person responds with a comment or another question. For example,

"Do you take drugs?" - Question

"No, I don't." - Answer

"Good." - Comment

Dominance relations between the participants can be established using these patterns. For example, the second person in an arching pattern has effectively taken control of the conversation if the first person answers his question. In chaining patterns, the first person retains control.

Patterned action sequences manifest themselves in an individual's behavior because they are useful in predicting or specifying future events and states.

The fact that an action sequence is usually carried to completion seems to prevent other

possible actions from occurring [Dennett 1969]. For instance, once one's hand is on the telephone receiver, almost no thoughts or actions will interrupt putting the receiver to the head and saying "Hello". Thus, actions are partially determined by the previously performed actions. Also, we note that recognition and interpretation of situational inputs will also be partially determined by previously performed actions. In our example from a conversation, the final statement "Well, it's not too interesting..." is interpreted as an answer to the original question "How do you like your work?" and not to "I was concerned that you were upset."

Because action patterns or sequences seem to run to completion once begun, an economy of information processing can result [Dennett 1969]. For example, when first learning to drive a car with a stick shift, one must perform each step of the gear-shifting patterned action individually, and verify the state of the world after each action. After years of practice, one needs only to initiate the action pattern. Since the action pattern has become automatic, one is free to think about other things while shifting gears. In the same way, the formation of linguistic actions into patterns optimizes information storage about dialog participation. Scragg [1975] described these two phases of action optimization as rote-oriented and goal-oriented actions.

Action patterns tend to overlap one another so that their component actions are multi-attributed. The following example is a portion of a psychiatric interview, with some sample action patterns diagrammed on the right:

"Where do you work?"	Question	
	|	
"I work at Sears."	Answer	
"How do you like it there?"	Fear	Question
	|	|
"Why do you want to know?"	Protect	Clarification question
		|
"Maybe it upsets you."	Topic intro	Clarification
	|	|
"It's ok. But I am upset."	New topic	Answer
	|	
"About what?"	Accept new topic	Question
		|
"Didn't I tell you already?"		Question
		|
"You mentioned something about bookies."		Answer

Note that one linguistic expression can perform actions at several levels. The patterns (Q - CQ - C - A) and (Topic intro-New topic-Accept new topic) overlap each other. For instance, the sentence "Maybe it upsets you" serves to clarify a previous question, retain control of the conversation, and introduce a new topic. Action patterns overlap and represent various levels of information. This fact extends the economics of possible information processing For example, Scragg [1975] pointed out that an action could be locally or globally modified to serve various purposes, e.g., walking as opposed to strolling. The patterns correspond to several levels of actions: actions controlling the conversation; actions responding to affect needs (Fear → Protect); actions determining the syntactic form of the output expressions; actions guiding the topic of the conversation. For instance, the response "Why do you want to know?" can be interpreted as a response to a threatening situation. The input "Didn't I tell you already?" is an attempt to keep control of the conversation.

The patterned action examples given above are from a third person's point of view. For our

purpose, it is necessary to implement patterned actions as rules for guiding the model's behavior in the first person. Each pattern can have a number of elements. Two types of elements in the patterns are: (1) actions that the model can perform, and (2) wait-actions [Scragg 1975] or recognitions of situations. (A wait-action is a step in a patterned action in which the model must wait for an agent in the environment to respond.) In our dialog example, the question from the interviewer "How do you like it there?" is not an action that the model can perform. The question is a condition or event in the world that should occur in order for the action pattern to fit. The interviewer's question must be performed by the interviewer; thus the action of the model is to wait, or even anticipate, the interviewer's next input expression. In general, elements of patterned actions are either events or states of the world that a person seeks to recognize in the world, or instructions that a person performs.

Psychopathological models currently deal primarily with defensive cognition. Defensive cognition consists of the strategies for manipulating beliefs and performing actions in accordance with those beliefs to minimize the overall occurrence and turbulence of negative affects. Defensive strategies are strongly dependent upon the notion of affects influencing other thought processing. One characteristic of these strategies is the biased interpretation of situations according to affect levels. For example, in the paranoid mode (according to the theory we are attempting to model), when the shame-humiliation affect rises to an extreme level, the person has a strong bias towards interpreting another person's actions as persecutory. Another characteristic of defensive strategies is that they are mainly unconscious. In paranoia, a person may attempt to find evidence that another person is persecuting him, but the use of this strategy for avoiding or reducing humiliation is unconscious. Defensive strategies also can have thematic responses to particular affects. For example, the fear affect generally triggers an attempt to protect the self; the anger affect tends to trigger an action of locating the cause of the anger in someone else's actions and contending with the person. In general, two sets of factors determine the recognition or interpretation of events and the selection of actions to be performed. The two determining factors are: the inner environment (the needs and desires of affect), and representations of the outer environment (beliefs and facts about the external world). (Simon [1969] described the relation of these two factors to the development of artificial systems.) If we consider these two factors both to be appropriate conditions for determining the use of patterned actions, then the strategies for defensive cognition become quite similar to patterned actions.

Section 4.1 Action Patterns

In the next three sections, we will describe a theoretical model of patterned actions in humans. Later in the chapter we will describe the computer simulation model which we based on the theoretical model.

The theoretical model of patterned actions will attempt to explain how actions are grouped into patterns and how these patterns are used in the recognition of situations and the performance of actions. The model will attempt to account for a particular type of actions: those actions that we consider automatic, unconscious, or not requiring deliberative planning to any great extent. These actions are rich in situational factors and affect but sparse in abstract thought.

The model's basic elements are: (1) data structures called patterns, (2) an activation mechanism that activates patterns when the patterns match a real-world situation, and (3) an interpreter that performs the actions specified by activated patterns.

In its simplest form, a pattern is a list, a data structure consisting of a one-dimensional set of elements that are linked. Each element of a pattern represents an event or state of the world. Elements are placed in a left-to-right sequence in a pattern according to the actual temporal order of the events'

occurrences in the real world, making the elements time-dependent. For example, in a gear-shifting pattern

(Disengage clutch → Shift gear lever → Re-engage clutch)

the elements correspond to the sequence's actual time ordering in the real world, with earliest events leftmost.

A more complex pattern may be more than one-dimensional by having a number of elements in the same position. The events corresponding to these elements are all supposed to occur simultaneously. For example, we may expand the previous pattern to include the events involving the throttle:

Disengage clutch & close throttle
|
Shift gear lever
|
Re-engage clutch & open throttle

In this case, the two starting events (Disengage clutch and Close throttle) occur simultaneously, as if conjoined.

Conceptually, elements of patterns are representations of events or states of the world (facts about the external environment), or conditions (affects) of the internal environment. This scheme allows one homogeneous representation of both factors influencing human behavior. We can interpret pattern elements as events that have already occurred or as actions or states that should be brought about, depending on how far along in the pattern the real world situation matches. For example, in the Clutch pattern example, if the environment matches (Disengage clutch), then the (Shift gear lever) element is an action to be performed, prescribed by the pattern. If the environment matches both (Disengage clutch) and (Shift gear lever), then (Shift gear lever) is an event that has already occurred. Thus, events in the patterns may be interpreted as situations to match or as actions to perform. Affects may be interpreted as conditions on the inner environment to match or as the affect response to the events in the pattern. Also, the elements of any one pattern are at a conceptually homogeneous level. The events that the elements represent are all of a similar nature. For example, the gear-shifting pattern would not contain elements from a pattern of conversational events, nor would it contain elements corresponding to specific muscular movements in the hands for shifting the gear lever. Such simultaneous pattern information on a different conceptual level would be included in a different pattern.

Section 4.2 Pattern Activation

Patterns may be active or inactive at any particular moment. A pattern is active when (part of) it matches the internal and/or external environment. The activation mechanism matches pattern elements against conditions in the environment as represented in the external sensory input and internal conditions. Patterns match from left to right (to correspond to temporal order). Each element of the pattern must match a condition; the last element matched must correspond to the most recent situation. When a pattern is activated, the remaining unmatched elements become available for further information processing. Inactive patterns are latent and unused in current processing. For example, the elements (Shift gear lever) and (Re-engage clutch) in the Clutch pattern can only be processed after the pattern has been activated and the (Disengage clutch) element has been matched.

Two forms of situation recognition are possible: recognition of new situations, and validation of on-going situations. A pattern activation represents a new situation being recognized. It also represents all the appropriate information being activated for the new situation. In our conversation example, the (question-clarification question) pattern provided a recognition of the local interaction as soon as the second element matched. The pattern then provided the appropriate information to recognize the next input as a clarification of the original question. Validation of on-going situations occurs when a next event is correctly anticipated, as in recognizing the clarification input, or when the next event is brought about through one's own actions, as in answering the question after clarification.

Patterns have initial conditions that determine whether they will be active or not. In general, initial conditions are the same inner and outer environmental factors that make up pattern elements. The necessity for making initial conditions special elements in patterns arises from having to keep a distinction between descriptive elements and prescriptive elements. In the (question-clarification question) pattern, we do not wish the fact that the other person asked a question to prescribe automatically a clarification question as the appropriate response. Instead, we need to add an initial condition that if the input was a question and if the question was not well understood, then a clarification is one of the possible appropriate actions to be taken. At that point, the two initial conditions plus the additional elements already matched in the pattern become the initial conditions to the next element of the pattern.

A generalized pattern in the theoretical model looks like:

(A1 & A2 &...& An → B1 & B2 &...& Bn →...→ J1 & J2 &...& Jn)

All of the Ni's are conditions. The conditions in any one pattern element are conjoined. The arrows indicate temporal order.

According to the theoretical model, the activation of patterns in humans proceeds continuously and in parallel. We can make an analogy between activation of patterns and a network of AND and OR logical gates, with a set of level inputs, the propagation of signals, and the final output of level signals. The level inputs in our theory are the inner and outer environment variables, almost at the level of sensory inputs to the system. The logical gates are the patterns, with an input connection between the environment variable and the element of a pattern that it matches. The final output signals are the set of next states in the patterns that are to occur. The propagation of signals and the activation of patterns can occur quickly because no graph searching or back-tracking need occur.

This combination of instant assimilation of new information and the simultaneous tendency to stick with the same pattern or interpretation of a situation that gives this scheme its useful qualities for psychological models. For example, the input expression "How are you?" can have (at least) two different interpretations. At the beginning of a conversation:

"Hello."
"How are you?"

the appropriate response is "Fine." After saying "My girl friend has an exotic disease," the input "How are you?" is undoubtedly more than a greeting. These interpretations can be distinguished by patterns of the following sort:

("How are you?" & Greetings → Standard greeting → Response="Fine")

("How are you?" & Nongreetings → Health question → Response="No infection")

We can make an analogy between the structure and activation of patterns as being similar to semantic nets. We can consider the system as consisting of a set of concepts or conditions on the inner and outer environments. For instance, "Humiliation high," "Input is a question," and "Conversation in greetings mode" would be concepts or nodes. The patterns are logical AND gates between concepts, so that one concept (e.g., "Humiliation high") would be linked to all of the patterns in which it was an element. (The time factor has been left out of this analogy.) Whenever an initial condition becomes true, if all the other initial conditions of a pattern were already true, the pattern would automatically be activated. Conversely, if a condition suddenly were invalidated, all the patterns in which it was an element would immediately be deactivated. Again, in our scheme, we reduce time-consuming matching processes by tying each condition to all the patterns of which it is a member. (Compilation of condition templates by network is more fully discussed in [Hayes-Roth 1975].)

Section 4.3 Pattern Interpretation

Once activated, patterns are processed by an interpreter. Each pattern has a pointer that specifies the element corresponding to the present. For example, the pattern (Question → Clarification question → Clarification → Answer) would have a pointer to the third element after the clarification question had been asked. The elements preceding the pointer represent facts; the elements succeeding the pointer represent expectations; the element pointed to represents an action to perform or situation to recognize. (All elements are specifically labelled as to whether they are actions that the self performs or wait actions that the person must recognize. This labelling scheme implies that there are two Clarification patterns, one for each role that the person can play.) The interpreter proceeds by: (1) examining all of the elements currently pointed to; (2) determining an appropriate action (which may be to wait and interpret); and (3) performing the action. The interpreter looks at only a cross section of the pattern elements in each cycle. The pattern pointers specify the pattern elements in the cross section. One might think of the interpreter stepping one element at a time across the patterns. Actually, the pattern pointers, not the interpreter, move across the patterns using the activation mechanism.

Note that the pointers do not move across all of the patterns. Only the pattern pointers whose information was used in determining the action change. In particular, a pattern whose action is performed by a subpattern would not be completed until the subpattern concluded its processing. For instance, a conversational pattern:

(Question → Answer → Comment on answer)

may have a subpattern performing part of its actions:

(Recognize input → Find answer → Say answer)

The pointer on the global pattern would point to Question during Recognize input, then would remain pointing to Answer during the last two elements of the subpattern.

The interpreter must determine a single action to be performed from all of the activated

patterns. This action may be a complex combination of several actions that are pointed to in patterns. For example, the linguistic output "Maybe you are upset about it" in the above Clarification example was a result of two component actions: (Clarification) and (Introduce new topic). To compose multilevel actions, the interpreter examines the current elements in the activated patterns and matches the elements with a small set of rules. From these actions, the interpreter selects a single action.

The actions that the interpreter executes are the only visible actions the system makes. The top-level loop of the system (assuming that the pattern activation is done continuously by an independent process) examines pattern elements, determines an action, and executes it (analogous to the fetch-decode-execute cycle of digital computers). We minimize conflicts between patterns by: (a) keeping the various levels of affects distinct in their interpretations of situations; (b) making the pattern elements as specific as possible in the conditions they represent; and (c) making certain categories of conditions mutually exclusive. If any conflicts arise concerning the correct action to take, one may either assume that all the remaining actions are appropriate, or that a state of confusion exists as to the best course of action. In any case, affect levels will change if no appropriate action is found, (e.g., shame is raised if the system detects confusion) until the affects are strong enough to dominate the system and insure some action.

Note that the interpreter can perform only one (multilevel) action or make one inference (if a pattern represents an inference) at a time, while the activation mechanism can recognize entire situations.

Section 4.4 Model

The remaining sections of this chapter will discuss the computer simulation model which is based on the theoretical model just described.

There are two difficult aspects to implementing our theoretical model: (a) the continuous and parallel activation of patterns; and (b) the difficulty, due to the relative autonomy of each pattern, of insuring that the right patterns will be active in a situation. As we shall explain later, the patterns are activated before the interpreter performs each action. The working memory is cleared of active patterns, except for those reactivated specifically by previous patterns and actions, and the entire set of patterns is tested for activation. A wealth of initial conditions and pattern elements which narrow the scope of most patterns achieves control over the autonomy of patterns. Unfortunately, this narrow scope also limits the patterns' generality.

We implemented the model in a Production System. Each production rule has the form:

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

where the Past conditions are mandatory and the other conditions are optional (indicated by brackets). There are two types of production rules: Action Patterns (APs) and Interpretation Patterns (IPs). An AP has the form:

(Past → Action : [Future])

An IP has the form:

(Past : Present)

Action Patterns specify some action that the system is to perform. APs are typical production rules. The Past conditions are initial conditions (left-hand side) of the rule. The Action element specifies an action for the interpreter to perform. The PS interpreter forms a conflict set from the APs and selects an action from the conflict set. Future conditions are set to true for the next interpreter cycle. Future conditions are only set to true if the Action in the AP is selected and completed successfully. APs model sequentially performed human actions. An example of an AP:

(THREAT → WITHDRAW)

Given that the interviewer has threatened the model-patient, this AP specifies an action of withdrawal.

IPs specify a situation that the model can recognize. The Past elements of an IP are the components of a situation to be recognized. The Present element is a set of conditions that are activated immediately whenever all of the Past conditions are true. The Present conditions can act as initial Past conditions to other rules. IPs model the parallel recognition (matching) of situations in humans. An example of an IP:

(STMT & FEARCHANGE : THREAT)

Given that the utterance was a statement and that FEAR increased at the same time, this rule interprets the situation as threatening to the model-patient.

Before examining the details of pattern elements, we shall describe the PS interpreter.

The five stages of the PS interpreter perform one cycle of activation and interpretation. The five stages are: Update, Activate, Match, Execute, and Book-keep. We shall give a brief description of each routine.

(1) Update - This procedure updates the appraisal and affect conditions. The physical state of the model's processing is assessed and the appraisal conditions updated (e.g., tiredness, cognitive overload). The affects are decremented slightly from the previous cycle to simulate a time decay of affect strength. The external inputs are tested, and any new information is noted in the corresponding appraisal variables. All of the conditions (internal and external environmental variables) that were set or are still true from the previous cycle are set true.

(2) Activate - This procedure activates all patterns that are true based on the conditions set by Update. Activation is accomplished by (a) testing all the conditions (stored on a list) for being true, (b) accessing the patterns for which the true conditions are pattern elements, and (c) testing whether all of the other elements in the pattern are true (i.e., contained in the "true condition" list). If all of the Past conditions are true, the pattern is activated (put into the "conflict set"). As each IP is activated, its Present conditions are immediately set to true. These conditions are also tested by Activate and may make additional Past conditions true. The activation process continues until all of the original conditions from Update and the additional conditions from Present elements of IPs have been tested.

Note - the activation mechanism in the Activate process is an exhaustive search through all of the patterns. This search is necessary because the IPs can add new true conditions during the Activation process. Also, there is no ordering of rules during Activation and no partial matching.

(3) Match - This procedure matches all the Action elements of the activated patterns (conflict set) with a set of action rules to combine the component actions into as few multiple attribute actions as possible. The multiple actions are ordered. Actions that serve as a response to high affect activation are placed before actions serving as low activation responses. Match chooses the first multiple action (first match) as being the most important action to perform.

(4) Execute - This procedure performs the (multilevel) action found by Match and determines which Future conditions to update for the next cycle.

(5) Book-keep - This procedure clears the working pool of active APs and IPs. It also determines which conditions have been set to be true at the start of the next cycle by Future pattern elements and resets all other conditions.

The syntax of the Past pattern elements is as follows: Members of the Past conditions are either LISP atoms (e.g., A) or lists of two elements (B C), where B and C are LISP atoms. The atom form A is true if the variable A is set to a non-NIL value. The list form (B C) is true if the value of the variable B is C.

Appraisal conditions and affects are two types of Past pattern elements. The appraisal conditions and affects are the facts that the model must interpret. Examples of the appraisal conditions are the interviewer's linguistic input, facts about the input, "the model is tired", "30 seconds have elapsed since the last input." Appraisal conditions represent all of the information from the model's interface to the external world. Other appraisal conditions are output conditions from IPs, e.g., THREAT. The affects are the six primary affects referred to previously. Each affect is represented by an atom that may have one of three values - LOW, MED, HIGH. For example, the condition (FEAR HIGH) would be true if the affect of fear were activated to a high level. These two sets of conditions are exogenous in the sense that any belief that the model forms about them will not change their values directly. However, patterns or situations that occur may indirectly modify the conditions.

The Present pattern elements are conditions set to true in the Activate process. The syntax is as follows: A Present condition may be either an atom A, a list with an affect and an increment (FEAR .10), or a list of two elements (B E). If the condition is an atom, the atom is set to T. If the condition is an affect-increment list, the affect is incremented by a factor (described in the next chapter) proportional to the increment. If the condition is a list, the expression E is EVAL'ed and the atom B set to the resulting value.

Present element represent interpretations of inputs. They are typically either affects or combinations of other inputs. For example, the IP

(INPUTF & (QUESTYPE STATEMENT) & (TALK I) : STMT)

generates the element STMT as a shorthand for the fact that there is an input, its type is STATEMENT, and it is the model's turn to speak. An example of an affect Present condition:

(CONFUSED : (SHAME .10))

This IP raises shame whenever the condition CONFUSED is true. CONFUSED is set whenever the MATCH process has more than one action in the conflict set.

The Action elements are LISP atoms that name actions. Match uses a set of action rules (called VPATs) to combine individual actions into multiple attribute actions. A typical VPAT is:

<< SHOW-ANGER & REPLY ; (ANGER-REPLY REPLY) >>

The form of a VPAT is:

<< ACTS ; FUNCTION >>

where ACTS are LISP atoms that name Actions and FUNCTION is an arbitrary LISP expression to be EVAL'ed. In the example above, SHOW-ANGER and REPLY are also elements of Actions in the conflict set; (ANGER-REPLY REPLY) is an action to reply with anger. The variable REPLY (on the right-hand side) is an argument to the function and contains a pointer to the semantic content to be expressed.

We use the Future pattern elements mainly as tags to link multiple step APs. An element may be an atom A or a list (B E). The atom A is set to T and the variable B set to the result of EVAL'ing E. The interpreter only sets the Future conditions to true if the corresponding Action was successful (returned a non-NIL value). Future conditions are set after the Bookkeep process has cleared all conditions in the model at the end of the cycle.

MPATs represent multiple step patterns, patterned actions which require more than one cycle to be executed. An MPATs general form is:

```
( ACOND (WCOND1) : (COND1) ACTION1
         (WCOND2) : (COND2) ACTION2 ... )
```

where the ACTIONs are the same as Actions in APs, the CONDs are the Past conditions in APs, and the WCONDs are wait conditions. All of the conditions and actions are optional. WCOND2 specifies that, after having performed ACTION1 successfully, the MPAT is to remain activated until WCOND2 is true. At the time, COND2 must be true in order to do ACTION2. If COND2 is not true when WCOND2 becomes true, the MPAT is deactivated. The ACOND is a condition that must be true throughout the activation of the MPAT. If ACOND becomes not true at any time, the MPAT will be deactivated.

The Clarification MPAT is:

```
(NIL (CONV)*(QUES) (INPUT)*(CLARIFQUES) (INPUT)*(CLARIF) ANSWER)
```

MPATs are conceptually equivalent to those in the theory, but have been broken up for ease of implementation. Each MPAT is broken into a series of APs. For example, the three-step pattern

```
(Recognize-input → Find-answer → Say-answer)
```

is broken into three single-step APs:

```
(Input → Recognize-input : TAG1)
(TAG1 & Recognized-input & Question → Find-answer : TAG2)
(TAG2 & Answer → Say-answer)
```

This decomposition into single-step patterns is done algorithmically by a separate program. The program inserts tags (in Future conditions) to link single-step patterns that originated from a single multi-step pattern.

MPATs can also have names indicating they are subroutines. For example, the AP

```
( STMT → REJOIN )
```

specifies an action of rejoining an input statement. The REJOIN action can either be a single-step action (LISP function) in a VPAT or it can be a multiple-step MPAT. For example, one of the model's methods to find a rejoinder is:

(REJOIN (POSAFF) :(INPUTF) VERIFY REACT GETRESP SAYRESP)

The MPAT specifies action to verify the input's truth content, react (emotionally) to the statement's truth, find a response, and say the response. GETRESP has further MPATs to get the next line in a story, agree with the interviewer's statement, or assert a general statement of belief on the subject. While this MPAT is being processed, the interpreter automatically keeps the original (STMT → REJOIN) AP activated.

Section 4.5 An example of the interpreter cycle

The following patterns are part of the processing for the input "I find you interesting." The general form is

Pn (Past : Present → Action : Future)

where the several conditions within each part (e.g., Past) are conjunctive.

P1 (INPUT & STMT & (INPUT e8020) : DINTEREST)

P2 (DINTEREST & STMT & (SHAME HIGH) : INSULT & (ANGER .10))

P3 (DINTEREST & STMT & (JOY HIGH) : COMPLIMENT & (JOY .10))

P4 (INPUT & INSULT → DEFEND)

The atom e8020 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. If these two conditions are true, The first pattern P1 asserts the belief DINTEREST. DINTEREST is true if the interviewer believes that the model-patient is interesting in some way. DINTEREST is a global belief which can be accessed by all other patterns. The belief is part of the 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 joy 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:

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

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. Immediately, the Present condition of P1 is set true, and the belief DINTEREST is set 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.

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 responses (e.g., DEFEND, WITHDRAW) are responded to before positive ones (e.g., PROBE, INTRO-TOPIC).

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

Book-keep - deactivates P1 thru P4, with the exception of P4 if DEFEND is performed by a multi-step pattern.

These five steps constitute a loop which continues indefinitely, the patterns determining what the program does at each step (i.e., the program does not have to await input to operate). Interesting conditions arise when no actions can be found, or when there is no external input to provide new information. Then the model must act on its own, striving to fulfill its own needs and desires by "entertaining" itself with new actions to attempt.

When Match finds no actions to execute, it sets a condition DEC which will be true in the next interpreter cycle. DEC can then be tested like any other condition in a pattern. If DEC is true and the model has not yet replied to an input, either the model cannot find an answer to a question or the model has no rejoinder to the statement. The model can then try another method to answer the question or assert that there is a lull in the conversation and start a new topic.

From a practical standpoint, APs allow a conversational action pattern such as (Question → Clarification question → Clarification → Answer) to be activated over a number of input-output pairs (as seen above in the MPAT example). At the same time, in the interval between producing output and recognizing the next input, the model can still activate other APs to perform inferences or evaluate the model's performance. If an attack comes in the midst of a clarification AP, the attack response can interrupt the previous AP by choosing an action with a high affect content. The attack can change the conditions in the system so that the clarification AP is not reactivated, and the AP no longer interprets the situation.

The action pattern formalism attempts to account for the direction of actions by internal and external factors. The environmental factors can instigate and influence actions and interpretations at the lowest possible level. We have defined external factors as the system's input from its interface with the world. But what are the internal factors that will insure the system's survival and enhancement? The next chapter describes how the affect mechanism is incorporated into the AP-IP structure.

Chapter 5. Affect

In order for an entity to survive, to interact with its environment to its own advantage, it must have a mechanism to measure its performance. That is, an entity must have a mechanism to measure the significance of immediate factors in both the inner and outer environment. This measure must be correlated with the environment's actual value to the entity, as measured by some impartial outside observer. This mechanism must match the interaction's complexity and subtlety.

In this chapter, we present a model of affects. First we describe a particular theory of human affect and highlight the motivational function of affect. Then, we describe how the theory is implemented in the simulation model and how the affect mechanism motivates the model's behavior.

Section 5.1 Theory of affect

Our model of affect (emotion) is derived from differential emotion theory [Izard 1971]. We extend this theory to include affect as motivation for other thought processing [Faught 1975]. According to the theory, affect is one of the five processes in a human personality. (The others processes are homeostatic, drive, cognitive, and motor). Affect provides the principal motivation for human behavior by invoking and directing other processes.

Affect is not a system; use of the word "system" would imply a set of processes that were relatively independent of the remaining personality processes. Affect is a theoretical concept which we will use to identify a particular set of internal structures and functions in human personality. These structures include states of the system, causes and effects of those states based on receptors, physiological components to the states, and resultant actions performed by the person. Affect does not motivate all action. The characteristics of affect's motivation are: (1) the action is due to some specialized state which reflects a measure of the person's quality of survival, and (2) that measure has a cognitive component. The resultant action from the specialized state must have been cognitively processed. Thus, reflex actions are not affect responses. However, affect can guide a person's processing decisions at extremely low levels. For example, Solley and Murphy [1960] point out that affects can alter the selection of incoming stimuli and the system's sensitivity to them. Erdelyi [1974] suggests that selectivity of information is not just in perception or response, but is in many processes. His examples are pupil fixation bias, pupil occlusion bias, pupil size bias, nervous accomodation bias, cognitive encoding bias, LTM search bias, STM rehearsal bias, and output bias.

The basic application of affect to human mental processing is as follows: First, specific conditions in the inner or outer environment (detected by perception or cognition) cause one of these special states to be assumed (affect activation), implying a value on the conditions [Arnold 1960]. The state then redirects processing (affect response), either through interpretations of situations or through actions to be performed. The conditions are valued positively or negatively, depending upon whether the condition enhances or detracts from the quality of the person's survival. A condition may be measured or valued based on its immediate effect on the person, or based on the person's past experience with the condition.

Section 5.2 Components of the theory

The affect process consists of: (1) several discrete affects, each having its own distinct motivational and phenomenological properties, (specifically, the distinct sets of phenomena that activate them and their motivational responses), (2) a pain-pleasure scale of the relative (un)desirability of each affect state, and (3) a mechanism for applying the affect state to the remainder of the system to invoke and direct other processing. Tomkins [1962] distinguished eight affects. The three positive affects are enjoyment-joy, interest-excitement, and surprise-startle. The five negative affects are fear-terror, anger-rage, distress-anguish, shame-humiliation, and contempt-disgust.

Affect has complex chemical, neurophysiological, neuromuscular, and phenomenological aspects. Affects are activated by perceptual activities such as perceiving an important object or situation in a person-environment interaction, or by cognitive activities such as memory or as imagination of previous and future events. Less common affect activations result from spontaneous motor, endocrine, or neuromuscular activity. Activation entails establishing patterns of electrochemical activity in the nervous system (through the limbic system) for a particular affect. These patterns initiate neuromuscular activity, primarily in facial activity and facial patterning, and secondarily in general organismic arousal.

When this activity is sensed at the phenomenological level, the activity becomes a conscious experience of a discrete, meaningful affect. The power of affect is in this experience: the experience is intrinsically rewarding for positive affects and punishing for negative ones. The experience is conscious in and of itself; affects do not require cognition in order to be sensed in consciousness. The distinction between drive and affect is in their access to consciousness. Drives (e.g., hunger, thirst) constitute needs of the system, but drives only motivate through affect. Drives send information about the person's physical needs to the affect process. Affect may amplify, attenuate, or ignore the information, thereby modulating the impact of the drive on consciousness. If the person has no conscious experience of the need, the need has no motivational influence. Thus affect can motivate without drive, but drive cannot motivate without affect.

Affect only indirectly influences the homeostatic and drive systems. The homeostatic system (comprising the cardiovascular and endocrine system, among others) is separate from the other systems and carries on its functions according to its own needs. The requirement of those needs is to attain or return to preferred states of the system in spite of repeated oscillations out of these states. Affect's only influence upon this system is from facial and body changes of affect responses. Drive system needs satisfy tissue demands for food, water, air, etc. Drive needs are periodic in nature and transmit their states to affect through affect activation. Because drive needs are only one class of affect activator, other activators may cause affect to ignore drives, to attenuate drives, or to amplify them, depending upon the other concurrent activators. Affect's response to the drive system is limited to directing conation and cognition to satisfy drive needs.

For a task of simulating human behavior, it is necessary and useful to model affect as it occurs in human personalities. To construct a program with a task to perform, it may be useful to implement an affect-like process representing the program's self-interests. The requirements of such a program typically entail finding and executing a "best" sequence of actions to perform a task. For the program to strive to do its best, it must perceive its relative success of performance. The relative success of performance must then take on importance and be significant to the system. The experience of this significance is an immediately rewarding or punishing experience, as in affect. Such experiences become synonymous with the program's self-interest. Using an affect-like process, motivation for all action would stem from this self-interest and would insure that the program executes all actions and makes all decisions for the purpose of completing the original task. If the program performs a task that is complex, we would expect the motivation of the program towards its task (and, more specifically, towards its self-interest) to be equally complex in guiding the entire system.

Section 5.3 Individual Affects

This section contains a discussion of the eight primary human affects and a set of hypotheses about their individual uses in humans.

Affects depend upon or are derived from two measures of the merits of an input to the person, or rather, two ends of a scale of beneficence to the person. We call these measures pain and pleasure. The only definition of a painful stimulus is that the person attempts to reduce or avoid it; the only definition of a pleasurable stimulus is that the person attempts to enhance or prolong it. We may attempt to further hypothesize that the physical mechanism of pain is an overload of sensory input so intense that the affect process automatically moves to reduces it, or that pleasure is the exercise of sensory channels to their maximum efficiency. But we are still left with the basic definition: The only effect on a person's information processing is the avoidance of pain or enhancement of pleasure. We make no attempt to account for the subjective experience of these states. There is no way we can objectively measure the subjective experience of pain in humans or any other organism. We can only assume that the person's subjective experience is similar to ours when he behaves in similar patterns and reports experiences which we take to be similar to our own [Apter 1970]. We define the distress affect to be painful and the enjoyment affect to be pleasurable. That is, the distress affect measures pain; pain activates the distress affect. Similarly, the enjoyment affect measures the pleasure a person is experiencing.

There are several sources of pain and pleasure. Drives have their input to the affect system through distress and enjoyment. Physical pain that is due to an unsatisfied drive activates distress. Physical pleasure activates enjoyment. Two types of mental or personality pain and pleasure also activate distress and enjoyment. An affiliation scale notes how much interaction the person has with other humans. Lack of affiliation activates distress which is interpreted as loneliness. Also, an esteem scale notes how proficient the person is at performing functions or obtaining wants. We postulate that the affiliation and esteem scales (in addition to physical pain-pleasure) also lead directly into distress and enjoyment.

The affect mechanism's strategy for negative affects (those which are painful) is feedback-oriented: whenever a negative affect is activated, the person attempts to reduce the conditions which caused it. For positive affects, the person's strategy is to not change the conditions and thereby enhance or prolong the positive affect or pleasure.

To summarize: The elements of the affect process are the eight primary affects, the set of neurophysiological inputs which activate them, and the affect responses. Each affect response is a physiological response, including chemical changes in the body and a facial response. But the particular affect activation levels are also available for sensing by the cognitive processes. The affects guide behavior by sensing the affect activation levels.

There has been some question in psychology as to the specific affect activators. Tomkins [1962] argued that all activators could be defined using a single principle: the density of neural firing or stimulation. He suggested that the various activators corresponded to different gradients in density. For example, fear is activated by a sharp increase in firings, anger by a high constant level of firings. The different gradients would be sensed by the affects and would activate them.

Izard [1971] maintains that an explanation of affect mechanism based on a single source is inadequate in accounting for the various information processing activities that activate affects. In addition to Tomkins' specific affect activators, Izard lists several types of mental processes which can initiate affects, including memory, perception, and motor activity. These processes would activate affects based on the individual's idiosyncratic experience and socialization. In general, the processes associate affects with mental constructs and are not specific to individual affects.

We shall define a scheme to account for affect activation from both specific measures of mental activity and general associations of affects with the content of mental constructs. The scheme is consistent with the theoretical model of patterned actions described earlier.

The eight affects in humans correspond to eight aspects of a person's relation to his environment that are important for the person's survival and prosperity. Each affect has a characteristic activator, a condition based on a particular measurement of mental activity performance or interaction with the environment. We define the substance or the essence of each affect by its characteristic activation condition:

Fear-Terror - anticipated pain

Anger-Rage - failure or thwarting of anticipated pleasure

Distress - prolonged pain

Contempt-Disgust - habitual pain

Interest - unanticipated element in the environment

Enjoyment-Joy - pleasure

Shame-Humiliation - lack of efficacy in action performance

Startle-Surprise - cognitive overload (e.g., overthrow of belief)

The eight activation conditions correspond to measures of a person's performance in the environment that are important for his survival. For example, a person must have some method of anticipating painful situations so that the impending pain can be avoided if possible. This anticipation itself must disrupt the person's mental activity to get his attention. For example, fear is unpleasant and therefore a negative affect; it stimulates avoidance of further anticipated pain. Anger enhances a person's continuity of actions. The person must anticipate pleasure so that he can continue a fruitful action and dislike its disruption. Distress allows a person to detect prolonged pain and deal with it. Contempt-disgust assigns negative value to habitual conditions or objects in the environment. Interest points out unexplained factors in the environment to be processed later so that new situations can be detected. Shame measures a person's performance and therefore provides motivation for self-improvement.

Section 5.4 Affect activation in humans

An affect can be activated or unactivated. If activated, it can be activated to any degree within a certain range. Two mechanisms, primary activation and secondary activation, lead to the affect's activation. Primary activation is based on a specific measure of mental activity. Primary activation is the result of a measurement of the environment and its immediate effect on the person. For example, the primary activation of fear is due to anticipation of future pain, as when one sits in the dentist's chair feeling the dentist's drill and anticipates pain. Secondary activation is a learned response to a situation. Secondary activation is the result of an association between an information processing structure and an affect. The structure represents a previous situation in which the affect occurred. For example, one can sit in one's own office thinking about the dentist's chair and feel

uncomfortable. There is an economy of processing obtained from not having to exert the anticipation mechanism for primary activation of fear, but instead having the secondary activation of fear directly tied to the memory of the situation.

The chemical or physiological properties of the activation of affects are not particularly important here, except to note that the stronger the activation, the more physical resources (e.g., adrenaline) are likely to be aroused for dealing with the situation. More important are the actual types of inputs which serve as primary activations (upon which secondary activations are dependent), and the types of responses which each affect characteristically produces.

Let us describe the primary and secondary activation of distress in humans. Distress is a measure of prolonged pain. That is, not only is there painful stimuli in the system, but there is a particular affect with its endocrinal response. For example, if a physically painful condition, such as hunger, exists in a person, he will notice it and attempt to deal with it. But if the hunger persists and is not satisfied, an additional painful experience, distress, which is greater than the sum of the combined hunger pain, will be added to the experience. This buildup extends over a period of time, so that responses based on minimal distress are attempted before those for maximal distress. This buildup thus alerts the person to unmet needs.

The secondary activation of distress is based on the association of a past event with another (primary or secondary) activation of distress. Distress is activated whenever mental activity accesses or processes the past event. For example, remembering a past poor stage performance can cause a person distress. Recognizing a familiar and noxious situation such as coming upon a fatal auto accident can activate distress.

In general, primary affect activation is used when the situation is new to a person; secondary activation is used in familiar situations.

Section 5.5 Affects in the simulation model

We now turn to the simulation model. With our model, we attempt to account for both primary and secondary activation. In addition, we want the modelled motivational mechanism to be the process that actually motivates the model. We shall describe the particular connection or condition of performance that activates each affect.

Given the AP-IP system described earlier, the primary affect activators measure critical aspects of the system. Specific procedures in the PS interpreter are responsible for primary affect activation. Each procedure measures an aspect of processing (usually only once per interpreter cycle) and activates the corresponding affect. Secondary activation is accomplished by including the affects in the right-hand sides of IPs, indicating that the affect should be activated whenever the rule is activated.

The model contains six of Tomkins eight affects (fear, anger, shame, distress, interest, and joy). Each affect is represented by a LISP atom with a numerical STRENGTH property whose value ranges from zero to ten. The LISP atom itself has three possible values, LOW, MED, and HIGH, depending upon the value of the STRENGTH property. The value is LOW if STRENGTH is 0-3, MED for 4-6, and HIGH for 7-10. The atom value is updated whenever the STRENGTH property is modified.

An affect's STRENGTH is modified in two ways. It is incremented by a procedure called

AFFMOD, or it decays over time by being decremented by a fixed amount. The input to AFFMOD is a list (NAME N) where NAME is the name of an affect and N is a numerical increment to modify the affect. The new value is a function of the old value and its associated increment. The values of the incremental variables range from 0 to 1. The new affect level is calculated according to the following formula:

NEW VALUE = OLD VALUE + (10 - OLD VALUE) X INCREMENTAL VALUE.

Thus, if the old FEAR value were 4 and the incremental value .25, the new FEAR value would be 8. The values of the affects approach 10 as a bound in such a way that early increments have the most impact.

Affects tend to subside when the stimuli that originally activated them disappear. In the model, the PS interpreter performs (during UPDATE) a calculation to simulate this natural decay in affect levels over time. Each affect is decremented a fixed amount.

Distress is a measure of prolonged pain. Primary activation is due to the interpreter measuring the activation levels of the other negative affects and incrementing distress in proportion to their sum. All of the primary activation procedures use the AFFMOD procedure.

The secondary activation of distress is based on the association of a past event with a previous activation of distress. The mechanism for this association is the inclusion of affects in the right-hand sides of APs as elements. The elements indicate that the affect's activation level should be raised whenever the pattern is activated. For example, the pattern

(CONFUSED : (DISTRESS .10))

increments distress whenever the CONFUSED condition is true. The pattern

(DMAFIA : (FEAR .20))

increments FEAR whenever the belief DMAFIA (that the interviewer is in the Mafia) is accessed. The affect increment (.20) is reduced an amount proportional to the strength of the belief (amount of evidence that the belief is true). The left-hand side on an IP is a condition or representation of environmental conditions; the right-hand side is an input to affect activation. With this mechanism, beliefs or interpretations of situations can (secondarily) activate affects.

With this mechanism, we can now describe the anticipatory mechanism used to activate fear and anger in the model. The anticipation mechanism arises from prediction of future events in the future elements of APs. The primary activation of fear is the anticipation of future pain or distress. Whenever the interpreter activates a multistage pattern that has as one of its later elements a secondary activation of distress, the interpreter detects the later distress element and instigates a primary activation of fear. If this occurs often enough, the secondary activation of fear may be incorporated into the AP by inserting an element to directly raise fear, without the anticipatory mechanism. Anger also uses the anticipatory mechanism. Whenever the interpreter activates an AP whose later stage includes a secondary activation of enjoyment, and that AP is deactivated without reaching the enjoyment element, then anger is activated.

Primary activation detects other aspects of system performance. When there is no AP to deal with an event, or when an AP is deactivated in the middle of its action, or when distress or enjoyment are activated unexpectedly (without being in the current APs as secondary activations) interest is activated. Enjoyment is activated whenever any one of the conditions defined to be pleasurable exists.

Shame-humiliation is activated whenever there is no AP to deal with a situation, or whenever an action does not bring about its expected result.

Note that the primary activation mechanism postulated here and Tomkins' mechanism based on neural firings do not contradict each other because they are at different levels of description. For example, our primary activator of fear is anticipation of pain, while Tomkins' is an increase in neural firing. An increase in density of neural firings may be the physical realization of pain anticipation.

In humans, secondary affect activation is dependent upon a previous situation for associating the appropriate affect to a mental construct. Whenever an affect is strongly activated, part of its response is to associate itself to the various process components that caused the activation. By "process components" we mean beliefs, goals, conditions - constructs such as elements of APs and IPs. Secondary activation is the mechanism for attaching significance or importance to process components. Since these associations are based in affect, they represent a particular process component's importance to process decisions - importance in terms of the self-interest of the person. For example, when a person infers a new belief that activates intense affect, the concept, event, condition, or rule of inference that led to the new belief will have an affect associated with it by means of the construction of a new association (IP) or the addition of the affect to an old one. For another example, when an action is performed to achieve a goal, the goal or action may have an affect associated with it. Later, when these process components are used in some action or interpretation, the secondary activation of their associated affects will point them out to the system as important.

Section 5.6 Affect response

Once an affect is activated, it responds by altering the processing in the system. The responses of affects, or the outputs of affects, can be primary or secondary. By primary we mean the affect has a characteristic response. For example, fear's primary response is for the system to withdraw. Anger's primary response is for the system to attack. Secondary responses are learned responses and are usually only characteristic to an individual or to a culture (as opposed to being characteristic to any person). For example, a person may learn to withdraw when angry. We shall return to the distinction later.

Using the AP-IP mechanism, affect guides the model's processing in two ways: (1) affects act as parameters to decisions, and (2) affects perform as an interrupt system. Affects are members of the left-hand sides of APs. Affects are like the system's other conditions in that they have their sensory inputs or neurophysiological inputs. Therefore, we can include affects into the AP mechanism as additional environmental conditions to be sensed and interpreted. Being in the left-hand sides of a pattern, an affect can determine an action or interpretation in the same manner as other left-hand side elements. For example, the input "I find you interesting" can be interpreted as a compliment when the joy affect is already activated as an insult if shame is activated.

P1 (INPUT & STMT & (INPUT e8020) : DINTEREST)

P2 (DINTEREST & STMT & (SHAME HIGH) : INSULT & (ANGER .10))

P3 (DINTEREST & STMT & (JOY HIGH) : COMPLIMENT & (JOY .10))

The atom e8020 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.

Situational interpretation by IPs perform the function of the appraiser subsystem of Averill, Opton, and Lazarus [1969]. They defined the three major components of the emotional response system as stimulus properties, appraiser subsystem, and response categories. The appraisal subsystem was a function of stimulus properties, motives, beliefs, and expected behavior. The appraiser served to reduce the amount of incoming information into an organizing concept such as threat.

We mentioned previously that a program's motivational guidance must be as complex as the cognitive processes that the program performs. This complexity is realized in the AP system by including affects in the left-hand sides of APs. For situation interpretations, motivational guidance takes the form of determining relevant information about a situation to be explored or a problem to be solved. The two sources of decision criteria for guidance are: (1) the tasks's global specification or situation and (2) information discovered by local processing as the task is performed. Cognitive processes use this information to guide which inference path to try next, which concept to elaborate, or which strategy to use for a certain task. Specifically, one cognitive task is to perceive a particular situation to obtain facts that are relevant to the program's purposes. Affects from the task's global specification determine which facts need the program's attention. Additionally, the association of affects with perceived objects or concepts provides amplification or attenuation of interest in the current local perception task as a source of locally-obtained information.

In goal-seeking behavior, affect responses help determine goals and actions and control action execution. The top-level goal which selects an action is a product of the currently activated affects and their positive or negative aspects. This goal is given its motivation by the current affects. When the program establishes subgoals of this goal, the subgoals are chosen according to their associated affects as being appropriate for the current affect situation. The subgoals then activate their own affects (from their right-hand sides) to add to the affect levels. The criteria for choosing the appropriate subgoals or actions are the left-hand sides of their corresponding APs. These left-hand sides include affect conditions. Finally, when the program monitors action execution, it may activate affects after perceiving the effects of those actions. The new activations may change the program's commitment to finishing the action by reducing the resources allocated if the program re-evaluates the original goal.

Affect responses also serve as a type of interrupt system. Affects act as a global system interpretation. For instance, the fear affect interprets the current situation as threatening. Using the AP activation mechanism described earlier, the fear affect's activation can interrupt current processing. For example, if the left-hand sides of patterns specify an action or interpretation based on a noneventful situation, the left-hand side conditions can change immediately whenever the system finds evidence of imminent danger. The existing situation interpretation changes entirely. All of the existing interpretations disappear from working memory (are deactivated), and APs to deal with the danger appear, interrupting the system in the duration of one interpreter cycle.

For example, in a paranoid mode, the question "Are you sick?" can be interpreted to imply "You are crazy" if the person is sensitive to humiliation at that point. The implication of being abnormal is then interpreted as an attack and dealt with as such. The following patterns show how we model this behavior:

((SHAME HIGH) & (INPUTF $SICK) : (INPUTF @SCRAZY))

($CRAZY : (SHAME .20))

$CRAZY is a pointer to the conceptualization (BE I CRAZY). $SICK points to (BE I SICK). The model processes the original input "Are you sick?" as a question until it finds the offending implication, at which point shame is raised. This immediately activates new patterns to deal with humiliation and insults. These new patterns re-interpret the situation. In addition, based on initial conditions of high

positive affect or low negative affect, some patterns may have been previously activated to promote cooperation with the other person. Those patterns are now deactivated and no longer figure in the situation interpretation since their initial conditions are no longer true.

Section 5.7 Issues and Implications

Besides serving as pattern elements, affects also serve (indirectly) as action selectors. The list of multiple actions is ordered according to the affect response content of the action. For example, the following is a list of some of the actions in the model. The list is ordered from high affect response to low.

(EXIT ENDCONVERSATION WITHDRAW DEFEND ATTACK DISTANCE
ACCUSE ANSWER REJOIN)

If both ATTACK and ANSWER are in the conflict set, the interpreter selects ATTACK to perform because it involves a stronger affect.

The distinction between primary and secondary affect responses is more one of description than one of actual processing differences. (Plutchik [1962] goes further and states that discrete responses for each of the primary affects are theoretical constructs only and must be inferred from behavior observation.) Secondary responses occur when affects are in the left-hand sides of APs. Secondary responses include most of the affect responses that a human uses. Primary responses are less tangible. For example, the primary response of fear is to withdraw. It is not so much that the system modulates the activity with a notion of withdrawing from the offensive or dangerous stimulus. Instead, it is that a withdrawal will force the offensive stimulus' deactivation and therefore the fear activator's removal. For anger, it is not so much that the anger response superimposes hostility on all actions, as the fact that an aggressive action may remove the obstruction to the anticipated pleasure that was disrupted. For the shame-humiliation affect, finding the cause of failure and locating it in the self or other will help to lead to corrective action to remove the cause for failure. Note that the person needs the concept of the self or a model of the self before such exploration makes sense. Similar factors hold for the remaining affects. The important factor is that the primary response or characteristic response to anger is hostility, not because of some direct information processing link between the anger level and the selection of actions which characterize hostility, but because hostile actions are the ones most likely to remove the activator of anger and are therefore more likely to be reinforced and remembered when anger is activated in a similar situation.

Secondary responses, however, are still oriented around a particular strategy for dealing with the situation (called sentiments by McDougall [1908]). The affect induces one basic strategy, such as denial, and then branches out into several variations of the strategy, such as denial of affect or denial of belief, depending upon the situation. These strategies may vary with affect activation levels. For example, mild fear activation may allow a person to accept the fear. Gross fear activation may induce fear denial because the acceptance of fear has led to more fear in the person's past experience.

One interesting secondary response is affect activation by other affects, e.g., fear of shame, shame of fear. If one is taught to be ashamed of fear, then the response to the fear affect's activation is to activate shame. Shame's response may be to deny the shame-invoking stimulus. If so, denial eventually becomes a secondary response of fear. When this happens, fear's primary response (to withdraw) will be overwhelmed by fear's secondary response, to invoke shame and a deny fear. Any combination of affect activations can be learned: anger of fear, fear of anger. Because they are based on past experience, the combinations are created as secondary responses.

Pragmatically, responses to affect activations can be almost completely dependent upon the affect or can be modulated by the external situation. Responses that are completely dependent upon the affect seem to be correlated with extreme levels of affect activation, such as shrinking in fear and lashing out in anger. These responses are dependent upon those rare situations in adult life in which milder responses are ineffective. More common is a response in which the external situation's features and the particular affect play a cooperative role in selecting an action appropriate to the situation which will at the same time cater to the affect's demands. Human seem to know lots of these combinations and have them at hand. They seem to know when lashing out is appropriate without taking time to think about the response's appropriateness.

This brings up the question of affect activation levels in responses. If we assume humans are born with a set of specific affect responses, e.g., crying for fear, screaming for anger, then at some point new strategies are added. In most people, the crying response to fear is still with them and is used only in extreme situations. This implies that either (1) the crying response has become associated only with higher fear activation levels and no longer all levels of fear or (2) the crying response was always limited to high fear activations. In both cases, new strategies would be introduced into moderate levels of fear activation to ward off the fearful stimulus before the higher level is reached. If the latter is the case, then the crying response in infants is a response to a high level of fear, and the situation (and subsequent memory of the experience) is more devastating than one might think. Perhaps the question could be resolved by measuring some external indicator of the fear activation level such as nervous system response and compare the crying response in infants to the crying response in adolescents or adults. In any case, the question should have relevance for developmental theories of affect influence in early childhood.

In a series of experiments, Schachter and Singer [1962] gave subjects an adrenaline injection and then either informed the subjects correctly of the expected effects, misinformed them, or left them uninformed. Schachter and Singer found that subjects who were either not expecting arousal or had been misinformed showed a marked emotional response to a social situation. This response mimicked a stooge who displayed either anger or enjoyment. Where the subjects knew what to expect, however, this knowledge seems to have completely inhibited their emotional responsiveness. Also, subjects given an adrenaline injection exhibited more affective response than those given a placebo injection when misinformed of the expected effects. Schachter [1966] concluded that "...an emotional state may be considered a function of a state of physiological arousal and of a cognition appropriate to this state of arousal." Further, "it is the cognition which determines whether the state of physiological arousal will be labeled 'anger', 'joy', or whatever."

Schachter's concept of emotional arousal corresponds to secondary activation in our model with one modification. In our model, the emotional state from secondary activation is due to the cognition causing the affect's activation. This activation is the physiological state mentioned by Schachter. This could occur in Schachter and Singer's experiment whether or not an adrenaline injection was given. We attribute the greater affect response from the adrenaline injection to a lowered threshhold of affect activation due to physiological conditions. We would model this phenomenon by modulating an affect's activation according to a measure of overall physiological arousal. Finally, we would explain the correctly informed subject's behavior by noting that cognitions leading to secondary activation can interfere with each other. The information that the subject will undergo arousal from the injection provides an explanation for the subsequent arousal. This explanation inhibits any action of finding the arousal's cause that would have otherwise occurred.

Finally, for affect to be effective at motivating an autonomous entity, affect must be complex. Including affects in both sides of APs allows for affect to match the complexity of the cognitive and conative processes it directs. The decision can be a global decision in strategy due to an affect response, or the performance measurement of a single action. Even these small measurements are tied to the global strategy of survival through the affect mechanism.

Chapter 6. Intensional Constructs

Intensionality in humans involves the application of the mind to a psychological object. The following properties characterize intensional constructs:

(a) Intensional constructs are symbolic; i.e., there is an internal (mental) construct that represents some external object, event, or property.

(b) A person can examine and report his beliefs, wants, and intentions [Dennett 1971]. A person's intensional constructs are reportable as internal events or mental states. (Nonintensional mental constructs can only be performed and not examined.) Hence, the person can communicate intensional constructs or build further models of them.

(c) Since their names are reportable, intensional constructs can be compared to their referents and updated for accuracy.

Intelligent systems make frequent use of intensional constructs to manipulate nonimmediate reality. Intensional constructs are not abstract types of symbols. Intensional constructs direct a person's actions by naming internal models of the outer environment (beliefs) and specifying a person's potential actions [Boden 1973] to alter the environment (intentions, actions). Boden [1972] defines an intensional action as a high level action that initially has to be consciously intended and carefully planned but can eventually be performed "automatically" and can directly control motor functions. In the previous chapters we described a system that has no intensional constructs but is still capable of responding appropriately to a number of situations using situation-action rules. However, without beliefs and intentions the model cannot represent the interviewer's long term characteristics, nor deal with situations in which its rules don't apply.

In this chapter we outline how intensional constructs can be modelled in our system. We indicate the need for intensional constructs as internal variables of nonimmediate reality. We examine theoretical problems of implementing intensional constructs including the problem of conation. (Conation is the direction of the body by the mind.) We then present the two components of an intensional construct: concept and token. A concept (for which there is no accessible label or name) can either be (1) an internal model that the system uses to discriminate a particular perceptual input or (2) an action to perform. Concepts are implicit in that a person may only access them by "executing" them (performing an action or discriminating a particular stimulus). Concepts embody a person's direct contact with the referent. A token (which labels a concept) is a reportable name. A token may be accessed by symbolic properties attached to it. An intensional construct has both components which may be accessed from each other. We describe how we implement these two components in the simulation model. We then give examples of how the model uses the two components in perception, planning, and error recovery after failed actions. Finally, we mention a number of global issues, including maximization of concept processing and comparing models to reality.

Note - we shall use the word "symbol" as little as possible because we need to distinguish between two of its meanings: (1) a symbol as a name or as a sign or token that represents something, and (2) just the token itself. Instead we will use the word "token" to refer to a linguistic name, with or without referent. (We use the word "token" similarly to the computational linguists' type-token distinction. "Token" refers to a linguistic name that the model can communicate. A concept is like a "type". We distinguish a token from an internal name in the model, e.g., a LISP atom "G0123". Since a token is simply a name, it can exist without a referent.) The word "symbol" will be used as a name with referent. In some sense, any internal construct in an information processing system is symbolic or is an internal representation. However, we will restrict the term symbol to linguistic names that label concepts.

Section 6.1 Background

In building psychological simulation models of intelligent behavior, a model-builder must eventually incorporate representations of nonimmediate reality. These representations correspond to possible or anticipated future situations as well as remembered past situations. Such internal representations are used for planning, reflective rehearsal, avoidance of undesirable events, aids or cues to perception, extensions to perception, and error recovery after failed actions. As an example: Suppose I am in a dentist's chair with my head back, and can't see my feet, and the dentist tells me there is a spider on my foot. I can now identify that crawling sensation on my leg based on the symbolic representation of the situation that the dentist provided. I then try to rid my foot of the spider by shaking my foot. I can ask the dentist for confirmation that the spider is no longer on my foot. If he confirms the spider's departure, I will be satisfied that I am no longer in danger based on the symbolic representation of the situation.

The important points to note are:

(1) A person can extend his perceptual abilities using symbolic representations; he is not limited to direct sensory input.

(2) Symbolic representations can activate affects (measures of emotional significance of situations) as easily as direct sensory inputs can.

(3) A person can anticipate situations which have little bearing on the current sensory situation and manipulate their representations. Action initiation is not totally dependent upon current sensory input but can be based on symbolic representations of nonimmediate situations. That is, the cognitive manipulation of intensional constructs can direct a person's conative activity. It is thought [Boden 1972] that as humans become adept at responding to their environment, more and more of their actions are based on these symbolic representations and less on immediate sensory input.

The need for and potential uses of these representations have been elaborated by Minsky [1975] in his frame theory and Schank and Abelson [1975] with their use of scripts. We will show how we use these representations in the AP formalism to enhance the situation-action paradigm.

Representations of nonimmediate reality are useful in certain situations. Are they necessary for intelligent behavior? The answer is Yes, according to Boden (1972). Intelligent behavior is characterized by showing efficient adaptation to the environment with respect to internal needs or desires. A person responds to environmental features to enhance his survival. Differential responses to particular environmental features are required to guide the choice and application of actions. To direct behavior in this way is termed perceptual discrimination.

To survive, a person must show efficient, appropriate adjustment to environmental obstacles and must be able to perceive goal state achievement to overcome such obstacles. This perception depends on an internal condition representing the external condition of overcoming the obstacle. Thus, the representation of nonimmediate reality is necessary for an efficient adaptation to the external environment.

A simpler, more intuitive reason for representing nonimmediate reality is the following: Without it, perception and planning would be totally dependent on the immediate situation for all cues. Often, there is incomplete sensory data from the environment and internal representations are needed to help fill in gaps and details. Using symbols enables a person to (a) perform actions that are independent of the present situation, (b) communicate symbolic knowledge, and (c) manipulate abstractions as a shortcut to manipulating reality.

Section 6.2 Problems of Intensionality

Common intensional constructs are beliefs, wants, and intentions. If we assume that such constructs are one of the factors causally responsible for a person's behavior, then we encounter several problems in modelling them.

The first problem concerns an internal model's completeness. How complete a model must an internal construct be before it is termed intensional? For example, a photocell can determine the presence or absence of light. It has an internal construct (current flow vs no current flow) which represents presence-absence data. But the internal construct seems too simple to be called a model. Another example is the belief "My father is a doctor" [Dennett 1969]. A child undoubtedly has a simpler model of doctors than an adult. If the child doesn't know what a doctor is, then does the statement "My father is a doctor" correspond to a belief? In the case of actions, we define an intensional action as an action that is called by name and has a body of information surrounding the action's concept specifying the action's use and probable outcome. A nonintensional or mechanical action is initiated by conditions in the internal and external environment with no reference to its name or consideration of its import. In general, there is no clear-cut definition about when an internal model becomes complete enough. Instead, we can only measure the degree of the internal model's completeness.

If we wish to construct a system that performs intensional actions (specified by name instead of by environmental condition), the problem of naming actions confronts us. Action (as part of an action pattern) has a temporal definition. An action has a beginning and end and can be subdivided in time. The question is, when can an action be subdivided into subactions, each with an (intensional) name of its own? The subdivision cannot continue indefinitely. If we agree that a person's intensional actions can motivate his behavior, then at some "bottom" level there must be a set of subactions which actually causes the action's physical realization. Otherwise an action would be only an abstract symbol and have no bearing on the person's conative activity. This is also true for interpreted computer languages. No matter how many levels of language interpretation there are in a computer, eventually there must be a machine code instruction that alters physical states in the machine.

The problem is determining how complex the bottom level actions are. One solution is to postulate a fixed set of primitive actions. The primitive actions that compose a complex action are called by their (intensional) names. There are problems with this formulation. First, no matter how we choose the primitive actions, a person can perform a fraction of one. Second, if a complex action comprises primitive intensional actions, then these component actions should be available to intensional recall and reflective rehearsal (introspection) [Dennett 1969]. If a golf swing is not a primitive, then a person should be able to report all of the primitive actions that comprise a golf swing as he is performing them. Finally, we assume that performing an intensional action requires a greater investment in processing time than a nonintensional action. If this is true, then it would be impossible to process the large number of primitive actions in a complex action such as playing a piano concerto.

Our solution to this problem is to define atomic and nonatomic actions. An atomic action is an action of arbitrary length or complexity that can only be performed from its beginning (e.g., the system cannot start performing the action from the action's midpoint). An atomic action cannot be subdivided. (Von Wright [1971] calls them "basic" actions.) We can make an analogy to compiled and interpreted computer functions. An atomic action is similar to a compiled function which can be performed but not internally examined. A nonatomic action is similar to an interpreted function which is made up of either compiled functions or other interpreted functions. The interpreted function, like the nonatomic action, can be internally examined and manipulated.

The third problem with implementing intensional constructs is their mechanical relation to a

system's action. The problem is to identify how an intention becomes more than just a goal to be discussed, how it actually generates action. Control must be passed to the appropriate action pattern to realize the intention's instigation of actions. A plan is not a physical entity, yet a plan can cause physical action [Boden 1972]. In computers, a plan causes actions by being an information pattern that directs the system's energy into particular paths. In our model we allow intensional constructs (beliefs, intentions) to be elements of left-hand and right-hand sides of action patterns. We will elaborate on this scheme later.

The fourth problem in modelling intensional constructs is the problem of conation. The question is how do the mind's information processes direct physical and mental actions? That is, who performs intensional actions? A mechanical action, such as avoidance of a fearful stimulus, can be caused by a stimulus-response process. The problem is explaining the causal mechanism behind a person's intensional actions.

The usual conative paradigm identifies the cause of intensional action A as another process B. However, if we claim that the person caused process B, we haven't reduced the problem at all. If instead we claim that process B is the person's want or desire to do process A, then it, too, is an intensional action, and we are in danger of an infinite regress. (Ryle [1949] points out and argues against a third explanation which he calls "the Ghost in the Machine". In this explanation, mind and matter exist separately. The mind is a "ghost" which resides in the body.)

Let us examine a different paradigm. Consider a system performing mechanical actions. At some point the system's actions fail and the system stops. We could construct an additional process of mechanical actions to locate the error and resume or restart the original mechanical actions. This process would use intensional constructs to hypothesize what the failed action was and specify the appropriate recovery actions. One example is a mechanical action to answer a summons, e.g. a doorbell. If the system starts an action to answer the doorbell, and then fails, the original sensory input of the doorbell is no longer available as a stimulus. The system must calculate that the summons-answering action is still appropriate. This calculation is a mechanical action, although it uses intensional constructs representing the original mechanical action. The calculation's result is to activate the name of an action to resume answering the doorbell. This name is another condition (in addition to affects and external conditions) to motivate a system's behavior.

The features of this paradigm are:

(1) All actions can be considered mechanical at some level.

(2) Mechanical actions can manipulate intensional constructs. In particular, mechanical actions can activate names of intensional actions.

(3) The system senses these activated intensional constructs like other (affect and environmental) conditions. The activation of the intensional construct is what causes intensional actions.

In this paradigm, an intensional action is the result of a mechanical action that is one level of abstraction higher than the intensional action. We shall defer the question of how many levels can be portrayed until a later chapter.

Section 6.3 Tokens vs Concepts

We have indicated the need for representations of nonimmediate reality. An intensional construct has two components - token and concept. We will use a redhot poker (Boden 1972) as an example. A child can learn the concept "hot" without the word "hot" by touching a redhot poker. The next time he sees a redhot poker he recalls his prior experience and avoids the poker. Avoidance stems from his internal model which the poker's sensory input triggers, not from the token (linguistic name) "hot". That is, there can be an internal representation (model) of "redhot poker" without tokens or names. A concept corresponds to an abstraction or class which a person can use to categorize objects in the real world. Also, the internal representation of a redhot poker need be neither red nor hot [Craik 1943]. The representation only needs to satisfy the perceptual discrimination requirements to include redhot pokers and nothing else. However, the internal model must somehow be analogous to its referent. "Hot" must be represented in a manner similar to other "hot" experiences to bring about similar behavior. Thus a concept refers to an internal representation of an unchanging state, event, object, or quality. Concepts are formed through a person's direct experience with the concept's referent. Concepts typically are based on some combination of environmental properties or predicates.

A token is the other component of an intensional construct. A token names or labels a concept. The phrase "redhot poker" is a token. Note that a token is artificial (forgoing onomatopoeia). A token's form has nothing to do with the concept's form that it represents.

Two examples of tokens in our model are (1) the belief that there is a lull in the conversation (LULL) and (2) the intention to tell about a problem the model-patient had with a bookie (PTELL). The corresponding concepts are (1) an IP by which the model recognizes a lull in the interview and (2) an AP with which the model executes an action to introduce the "bookie" topic into the conversation.

The distinction between token and concept lies in their usage. The system can access and manipulate tokens directly as names, without the corresponding concept being activated. A token can be a part of manipulatable links between other tokens. A concept associates the intensional construct to sensory input predicates ("reality"); a token associates the construct to other names. One problem in AI programs has been a confusion of the two components. In many programs there is no distinction between the concept "coldness" (an internal representation to recognize coldness) and processing the token "cold". Also, many tokens have no associated concepts and are therefore just place-holders. It is thought that concepts are usually formed first through experience, with tokens later attached to them. Fodor [1976] supports this assumption and states that "...organisms capable of learning a language must have prior access to some representational system in which (the semantic properties of its predicates) can be expressed." Of course, the semantic predicate or concept may be acquired in a number of ways [Winograd 1975], including modification of a similar concept, linking two component concepts, or direct sensory experience.

Both concepts and tokens are representations of external reality or observed internal behavior. An intensional construct represents nonimmediate reality only when the system can invoke the construct by means other than its conceptual sensory input (concept), implying through its token. In general, the concepts act like a stimulus recognition in a stimulus-response paradigm when used as interpretations of sensory input. However, when invoked by its corresponding token, a construct is a representation (internal model) of reality that may or may not exist in the current situation. Also, tokens may be linked to other tokens as models of reality. Intensional processing of nonimmediate reality occurs when tokens are used, either as links to other tokens or as handles on concepts.

Section 6.4 Representation of Concepts, Tokens, and Links

There are two types of patterns: action patterns (APs) and interpretation patterns (IPs). Action patterns generally specify actions to be performed in response to some situation or particular system state. Interpretation patterns put order into the mass of input data that the system deals with by interpreting large sets of input data and replacing the data with simplified sets ("chunking"). We use IPs to represent both concepts and tokens, although in different manners.

An IP represents a stereotype of the system's experience with some external or internal reality. All IPs are thus concepts, although some IPs may be too specific to be useful as a generalization of a real-world property. For example, the IP for a poker would be the concept for poker; the IP for recognizing a best friend's anger may be specific to the best friend's typical expression of anger. Note that concepts are still stereotypes. The concept for a poker encompasses any poker. The concept for a best friend's anger encompasses any instance of that person's anger. Like all IPs, concepts may be part of larger experiences and may have affects as their antecedents and/or consequents.

We also represent tokens as IPs. However, the IP has one extreme quality: The IP's antecedent represents the acoustic or visual occurrence of the token's name. The name is linguistic in origin and instantiation. IPs have a fleeting nature due to their dependence upon initial conditions to keep them activated. A token must rely on a different mechanism if it is to be independent of environmental conditions. A token's linguistic nature enables the model to rehearse the token in short term memory (STM) to provide this independence. (Note - we give short term memory only a cursory model in this work. Its only function is to enable tokens to remain active for the duration of one action execution.) Thus, tokens are independent of the immediate external environment. They do not rely on direct conceptual sensory input for their manipulation or activation. A token IP is linked to its corresponding concept IP: The token IP's output is an alternate activating antecedent of the concept IP. We represent tokens by LISP atoms in the model.

We use two types of intensional constructs in the model: beliefs and intentions. A belief in the model is a judgement about the truth of an internal representation. In the model, we use beliefs to represent the interviewer's long term properties, such as his abilities, interests, intentions, traits, and future actions. Knowing this information, the model can judge future actions for their possible success. In addition, beliefs represent the interview's state, the model's success at achieving its goals, and the outcome of the model's actions. Intentions represent states that the model attempts to bring about. Two examples are introducing problem areas that the model wants to discuss and asking confirmation for the model's beliefs. In general, we use intentions to give substance to the model's needs and desires by establishing goals for subsequent actions.

We can link tokens for various purposes. In the model, we use links between tokens for inference rules and explicit properties between tokens. For example, the connection between inference rule antecedents is a link. A belief is connected by a link to its property of evidence that the belief is true. A property between beliefs or names, such as "IS-A" in "A dog is a mammal" is also a link.

We represent links in the model in two ways:

(type 1) as explicit properties on tokens in a semantic net ("interpreted"). Examples are belief strengths and the relation between elements of propositional information.

(type 2) as an Interpretation Pattern ("compiled"). Examples are "compiled" inferences in which the inference rule antecedents are the left-hand side elements of an IP. The consequent of such an IP is a belief to be concluded as true.

(Type 1 links are the "interpreted" rules of Newell and Simon [1972]. Type 2 links are their "assimilated" rules.)

The model can execute "compiled" links (type 2) in some sense. The Update procedure in the pattern interpreter matches their component IP nodes. The Activate procedure traces their links. "Interpreted" links (type 1), on the other hand, require relatively more time to process. The interpreter needs the Execute procedure to perform a special action to follow this type of link from one node to another. This requires one interpreter cycle to occur because it is a sequentially-performed action. However, the system can access interpreted links to report them or directly manipulate them through other actions performed by Execute. The system cannot examine or manipulate compiled links (or any APs or IPs). The system may only execute compiled links if it can determine the appropriate conditions to set up in the environment. The system then can observe the outcome and inferentially attempt to reconstruct an abstract pattern representation.

The processing interaction between concepts and tokens is relatively straightforward at the lowest level. As mentioned above, IPs can be either active or inactive. The system activates a concept IP when the system detects the concept's associated antecedents in sensory inputs. Alternatively, the system activates a concept IP when the concept's corresponding token IP is activated. For example, the system can instigate actions based on a lull condition either by (1) detecting a lull with an IP or (2) activating the token LULL as the product of an inference or of new symbolic information from the interviewer (e.g., "I think the conversation is at a lull.")

Token IPs, as general IPs, can also be active or inactive. They must be sustained in a different manner from concept IPs since token IPs have no concept input antecedents. Instead, token IPs have their particular name and associated properties as their antecedents. The concept IP may activate the token IP, e.g., naming a bird with the word "bird". In our model, the concept IP which detects a lull may initiate an AP response to lulls or it may activate the token LULL. In the former case, the action may have no residual traces of its activating condition. In the latter case, the LULL token will be rehearsed and can be manipulated even after its activating concept condition is no longer true.

Section 6.5 Uses of Intensional constructs

The interaction of token IPs and concept IPs can be seen in the processing of beliefs, inferences on those beliefs, and intentions in the model.

One of the model's goals is to get help. In order to best seek help, the model must make inferences about the interviewer's abilities, interests, intentions, traits, and probable future actions. The model must also evaluate the outcome of its goal-achievement attempts so that it may better direct itself in its pursuits.

Secondly, the model must avoid threatening situations arising from threats of either mental or physical harm. For this the model must compare the interviewer and the interview to its concept of typically expected behavior. The model must then judge the interview as normal or abnormal, i.e., predictable and nonthreatening or unpredictable and potentially threatening.

Finally, in interacting with a psychiatrist, the model is sometimes called on to examine and evaluate its own behavior. For this task the model must examine its own behavior in the interview and evaluate the psychiatrist's opinion before making a judgement about its own behavior.

The belief system contains approximately 80 beliefs. Beliefs refer only to topic areas in which there can be evidence to change the belief in the course of one interview. Such areas refer to the interviewer, the interview, and the current state and intentions of the self, and not to relatively unchanging facts about the world or the model's own past history. Beliefs begin with default truth values indicating the self's assumptions prior to the interview. Truth values change during the interview's course.

We represent a belief in the model by a token (LISP atom) with the property of TR whose value corresponds to the amount of evidence that the belief is true. (The corresponding concept is an IP which is activated when the concept is recognized in the current situation.) The truth value ranges from 0 to 10, 0 indicating no information, 10 indicating enough evidence to conclude that the belief is true. Inferences can conclude that a belief is true (truth value TR set equal to 10) or add to a belief's truth value (truth value TR incremented by a positive integer). The negation of a belief is an entirely separate belief with a truth value from 0 to 10. For example, the following four beliefs are in the system:

DDHELP - the interviewer desires 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

This permits the model to conclude as true the belief *DDHELP independently of the belief DDHARM. Also, competing evidence can accumulate for both DDHELP and *DDHELP without having evidence for one cancel evidence for the other. Contradictions may arise if enough evidence accumulates to infer both the belief and its opposite. Contradictions may be noted by raising the interest or shame affect.

There are approximately 130 rules of inference in the model. Inferences correspond to the ability to draw new conclusions about new situations. In our simulation, inferences make possible the evaluation of the interviewer and the interview, examination of the self's actions, and prediction as to the future behavior of the interviewer. They invoke no actions themselves. Instead, other patterns in the model use their results as for making cognitive evaluations about the world.

We implement rules of inference in the model with IPs. The rule's antecedents are elements of the IP's left-hand sides. The consequent is a belief to be concluded as true or a belief with an increment to be added to its truth value. All inferences in our system are composed of compiled links (i.e., constructed from IPs).

A rule of inference in the model has a list of antecedents and a consequent. The antecedents are each tested for truth and the resulting logical values from the antecedents are conjoined. An antecedent may be: (1) a belief, which is true if its truth value is true (i.e., equal to 10), (2) a NOT-condition predicated on a belief, in which case the antecedent is true if the truth value of the belief is not yet true (i.e., not yet equal to 10, the maximum), or (3) a normal Past element condition. A consequent may be either a belief with a truth value of 10, in which case that belief is concluded as true, or a belief with an incremental truth value, in which case the increment is added to the truth value for that belief.

As with other APs and IPs, affects can be elements of either sides of inferences. In particular, an affect condition may be an antecedent in an inference rule, for example

((FEAR HIGH) & DDHARM : DGANGSTER)

This rule implies that the fact that the doctor desires to harm the model-patient is evidence enough, under high fear conditions, to conclude he is a gangster. An intention for revenge may be made during extreme anger:

((ANGER HIGH) : PHARM)

The system also sets new affect conditions when an inference rule or a consequent belief is set true. In the model, affects in the right hand sides of IPs dictate these new affect settings. Whenever a belief is concluded as true, the system sets any associated affects. For example:

(DMAFIA : (FEAR .40))

These affect conditions allow the system's internal needs or state to influence the intensional calculations of inferences.

Beliefs are also used for perception extension or perceptual cues. Perceptual input data is often incomplete in natural language processing. Often, the system has enough data from the input to activate the concept IPs which are antecedents to an inference rule. In this case, the system can use the inferential links from the activated concept IP antecedents to the corresponding token IP consequent belief and then to other associated token IPs beliefs. These associated beliefs act as cues or contexts to the perceptual routines by suggesting other concept IPs for matching to the sensory input data. For example, the inference

(DNBELIEVE & (SHAME HIGH) : DBABNORMAL)

states that when the shame affect is high, the belief DNBELIEVE (that the interviewer does not believe the model-patient) implies that the belief DABNORMAL (that the interviewer believes the model-patient to be abnormal) can be concluded as true. In the inference

(DBABNORMAL & (INTOPIC MENTAL) : (ANGER HIGH))

the belief DBABNORMAL can interpret subsequent inputs whose topic is the model-patients's mental condition as insults by raising anger. The belief DBABNORMAL acts as a contextual discrimination cue for the condition (INTOPIC MENTAL).

We use intentions in our system to direct processing with explicitly represented goals (as opposed to implicit goals in situation-determined actions). There are six explicit intentions in our simulation. Intentions represent action patterns that the system can perform to satisfy its needs. The action patterns in our current simulation include only linguistic actions that can be performed in an interview situation and are therefore specific to the task of interview participation. Intentions are represented in a manner similar to beliefs - a data structure with a property STRENGTH indicating the intention's current strength. STRENGTH ranges from 0 to 10. An intention becomes viable when its strength reaches a threshold of 9. An intention's strength is modified by inference rules.

Intentions represent situations which the model-patient believes to be advantageous (as the result of some prior experience) and wishes to bring about. Intentions are set to true by inference rules whenever the model is not engaged in responding to an input or dealing with some extremely distressing situation. The inferences that set intentions usually occur while the model is waiting for the next input from the interviewer.

Once an intention is set true, it becomes a pattern element like any other environmental condition. Two results may occur:

(1) The system treats the intention like an action and executes the corresponding action. In this case, the action may still only occur when the situation is appropriate, as the following example shows:

(LULL & PINTRO : INTROTOPIC)

The intention in this case reduces to an intensionally specified action.

(2) The system treats the intention like a belief or abstract representation to be symbolically manipulated. The system may then make inferences about the appropriate method to accomplish the goal and establish subgoals (rudimentary planning). The system may treat these subgoals in turn as intensional actions or abstract goals. For example, the inferences

(DDHELP & PTELL : PHELP & INTROMAFIA)

can establish the subgoal PHELP (to solve some problems the model-patient had with gangsters) and the intensional action INTROMAFIA (to introduce the "Mafia" topic) from the goal PTELL (to tell about his problems).

The system may also use intentions to restart failed actions. A patterned action to introduce a new topic may be activated upon detecting a lull in the conversation. A different action may interrupt the topic-introduction AP, for example by the interviewer continuing to mention a previous topic. The no-longer-detected lull may be recalled by remembering the model's previous output or by other inferential processing. The important point is that the lull condition is no longer available in sensory input to activate the concept IP or AP for introducing the new topic. If the system is to restart the routine, it must restart the routine through a token IP of the lull, for instance the belief LULL. The system then establishes an explicit intention to introduce the new topic.

Section 6.6 Global Issues

To summarize the preceding, we offer a system which performs all action with action patterns (APs) by activating the APs and then interpreting their included actions. The two types of data structures are action patterns and interpretation patterns (IPs). IPs are used for situation recognition. As concepts, IPs become the experiential representation of the concept they are used to recognize. IPs are also used as tokens or names of concepts. As names, IPs are used for recognizing the particular visual symbol name that they represent. The system accomplishes most of its processing by recognizing certain situations and responding to them with APs, in a situation-response manner. Note that this includes high-level processing. For example, there may be a special AP for adding one and one since it is such a common operation. Tokens are used to represent abstract qualities of the environment (beliefs) or future situations to obtain (intentions). Finally, if the automatic AP and IP processing fails for a situation, the system uses symbolic processing with token IPs for realigning the system with respect to the environment or revising the internal representations.

One efficiency principle for the system should be the maximal use of compiled concept IP links and the minimal use of interpreted token links. APs and concept IPs are specializations and represent knowing what to do with a familiar situation. The system should do as little processing as possible (according to technological principles) and should make maximal use of its resources.

Therefore, links between token IPs will soon become concept IPs to save the processing effort of activating the included token IPs and following the link. For example, we could imagine some future model being taught a symbolic link between a situation and action, for instance introducing a new topic during lulls in a conversation. At first the model would have to explicitly manipulate the tokens for the belief LULL and the intention INTROTOPIC. Later, after the model had performed the routine in enough similar situations, we could imagine the model developing an action pattern to detect lull conditions in the pattern's left-hand side and specify topic introduction in its right-hand side. The system would also compile basic symbol manipulating tasks into AP routines. Some of these routines might be: following an explicit symbolic link from one token IP to another, determining what situation the model is in (by activating a token IP which represents the type of situation), and determining what the outcome for a given action is likely to be. The APs will embody a set of basic processes to manipulate tokens.

The difference between the two types of links is useful in showing how a system could acquire information symbolically. The difference is between explicit, accessible, manipulatable links (as in a semantic net which the model can directly manipulate) and a link between IP elements. We hypothesize a pattern-creating process that will form new patterns (APs and IPs) representing repeated situations, for example a situation-action rule. (We have not modelled this hypothesized process in this work. We assume that the model has already compiled all inference rules. Becker [1973] attempted to explain the incorporation and subsequent generalization of "schema", which are similar to APs.) Interpreted and compiled inference rules are useful because a system could learn the interpreted rules directly by storing the rules in a semantic net. The system could follow the net links explicitly. The system would soon compile the interpreted inference form, if used often enough, into IPs for faster processing.

Another principle involves using internal representations and sensory input data in perception. As a system gains more confidence with the dual use of concept IPs and token IPs, the system will place more reliance upon the concepts. Actions will be based more upon IPs than on sensory input. Internal representations will be suggesting all of the hypotheses for situation interpretations and thereby guide processing in their own directions. This leads to a greater probability of confirming the suggested hypotheses and the hypotheses' later repeated use. Relying on the internal models results in an an information processing economy, but may lead to overlooking new changed situations and delaying adaptations to them.

Representing intensional constructs as tokens and concepts in the model has implications for the ease of learning new instances from another's teaching. A person can learn new tokens by the repeated use and rehearsal of either acoustic or visual form. He can learn new concepts by being shown numerous examples of situations in which the concept occurs. If the person had previously formed the concept, then he can be easily identify the concept internally and tag it with its appropriate token. If the concept was not already in existence, then repeated exposure to it, and simultaneous repetition of its token, should hasten the concept's formation and naming. Also, the token allows the pupil to guess at identification of novel objects and receive confirmation of his hypotheses. Further, a new situation-action rule can be taught by breaking the rule into its component situation concept and its associated token and the action concept and its token. Then the situation and action tokens can be linked symbolically and the action rehearsed until it is learned.

Another issue is interfacing specialized knowledge of common situations and situations that a system needs to respond to efficiently (e.g., danger) with general knowledge that the system can use in innovative adaptive ways. We can incorporate specialized knowledge into APs and concept IPs which will be specific to the situations needed. The same information could be processed by the token IPs of the concepts involved but only after having activated the token IPs and the links. Token processing is assumed to be slow (compared to pattern activation) since the process is a linguistic operation and

would require several interpreter cycles. The APs and IPs incorporating specialized knowledge deal with situations first. If these patterns fail, tokens can be used to apply generalized knowledge.

Finally, the intensional quality of concepts and tokens necessitates comparing them to reality to maintain their status as models of the external world. There are two forms of model - concept IPs and links between token IPs. A token IP by itself is insufficient to represent; it is simply a place-holder. When the system attaches properties, it either forms a concept IP to hold the property attachments or a link between token IPs. The link between tokens creates a corresponding concept IP if used often enough. The two types of model structures to be compared to reality are (1) concept IPs and (2) links between token IPs which have not become concepts.

The difference between these two types of models is in their accessibility to manipulation. The system can directly manipulate links between tokens by symbolic processes. Concept IPs can only be indirectly changed. (Of course there is a tradeoff - concept IPs are processed completely in the IP activation mechanism and require fewer resources.) Since a concept is automatically formed by observation and rehearsal, the only way a concept can be modified is by exposure to instances of the concept in the environment and emphasis on certain weaknesses in the concept (the weaknesses are the factors that need to be changed). The emphasis must come from tokens, since the original concept formation did not include the desired new factor. The tokens activate their corresponding concept IPs which then are emphasized and have a greater probability of being included in the revised concept AP.

Chapter 7. Meta-systems, Self-motivating systems, and Paranoia

In the preceding chapters, we outlined three essentially different aspects of motivational and intensional behavior: patterned actions, affects, and symbolic manipulation. We will now consider how these three aspects interact in larger issues. The issues: (a) the construction of meta systems to observe, interpret, and correct errors in their corresponding object systems, (b) the construction of self-motivating systems and their requirements, and (c) the elaboration of the theory of paranoia that we set out to model.

Section 7.1 Meta-systems

Section 7.1.1 Introduction

A particularly elusive task in AI has been implementing a program that models itself and its interaction with the environment. The twin problems of infinite recursion ("I know that I know that...") and lack of direction to the program's behavior (once the program knows, what is it supposed to do) make straightforward modelling difficult. However, if the program is to recover from its own failures, it must have some idea of the task it was performing when it failed and of its usual capabilities and expected outcomes of its actions for the task. Similarly, if the program is to deal with incompletely specified situations, it must have a grasp of the relevant features in typical situations. It uses these features to analyze its current situation and decide what is important to accomplish. For these tasks, an internal model of the program's own interaction with its environment is necessary.

We shall describe our efforts at implementing such a model. Using the simulation model of cognitive and motivational processes described in the previous chapters as a base, we added processes to determine three factors: (1) what action was taking place in the program and in other participants in the situation; (2) what actions the program itself determined to be most desirable; and (3) what action the program could perform in the particular situation and what the likely consequences would be. We shall describe these processes, and then describe what additional capabilities the processes make possible. Finally, we shall discuss the extent to which the program can make use of these processes.

To summarize, the object program is a feedback-oriented system that reacts to situations by performing the action specified by APs and IPs. It is guided by an affect system that attempts to avoid negatively-valued situations and enhance positively-valued ones.

Section 7.1.2 Description of the Meta-level

To this base we added a number of meta representations that explicitly specify the state of the environment. The representations are designed to facilitate answering questions about the system's own processing and the environment:

- What is the system doing and what has it been doing?

- What is the interviewer doing?

- What does the system want to do - i.e., out of all states that are obtainable from the current state, which are the most desirable as measured by affects?

- How could the system get there - what actions could the system perform to manipulate the environment to obtain the desired state?

- Is the system in control of itself? This is partially measured by how successful the system has been recently. It is useful for determining what level of risk is allowable and what certainty to put on perceptions.

Beliefs which answer these questions are represented in the same form as the remaining beliefs in the system. From these beliefs, the system infers the necessary information to deduce what actions it was performing and what actions it should attempt to perform. These beliefs represent action at one level higher in abstraction from APs in the base program. The belief representing the degree to which the program is in control of itself is necessary to marshall more resources or reduce risk when performance is poor.

A number of processes manipulate these representations to infer their respective beliefs. The processes can return two values for each belief: (1) a definite answer (2) or an indication that the program cannot determine an answer. If the system contained another representation level, a failure of one of these routines could invoke an action with a higher abstraction level and be processed similarly to this first level. (This potential will be discussed later.) The current program simply notes the lack of ability to determine an answer and decrements the level of control that the program believes it has over itself.

The meta level processes are called only on events which disturb the system. As long as the system is successful at performing appropriate actions (i.e., affects levels are optimal and actions are being performed), then the meta processes are not needed and are not used. The meta processes are invoked whenever a high negative affect condition or action failure occurs, or when the system is asked directly for an opinion or for information on its internal state (e.g., "What do you want to do right now?"). The meta processes are embodied in an action pattern (AP) (represented in the same formalism as all other actions in the system). This AP carries out the following individual actions:

- Determines what the system has been doing.

- Determines what the system is doing.

- Determines what the interviewer is doing.

The previous three make up the current situation.

- Determines whether the system finds the current situation desirable, i.e., whether the system likes the current situation according to affect measures of the situation.

- If the current situation is desirable, the system sets an intention to continue it.

- If the situation is undesirable, the system determines which states are possible to obtain, chooses the most desirable one, and sets an intention to perform an action leading to the selected state.

(The action pattern is actually broken into several small steps to make use of partial information if it is available. Each step is only used if its corresponding belief has not yet been determined.)

The end result of performing this action pattern is typically an intention to start an action. This intention is a member of the normal object level APs; if the situation is appropriate to the action

and no higher priority action is being performed, the intention will initiate an action. Note that the meta level routines do not have direct control over the object program's processing; their conative control consists of activating beliefs and intentions, relying on AP and IPs to detect the beliefs and intentions and initiate the corresponding action.

Based on these processes, the system can answer questions about its beliefs, such as what situation it is in, what situation it would like to be in, and what it is attempting to bring about. The system can also make a simple recovery from failed actions, although only by determining what the grossest action level was and attempting to restart the action. The system can perform in an incompletely specified environment by determining the state of the environment to the best of its ability, subject to its estimate of its control over the situation. It then acts on that environment with the assumption that the unknown is consistent with the known. The meta level processes can also influence the affects. If the system is successful at attempting to satisfy an intention, enjoyment (corresponding to high self esteem) is raised. Similarly, if its action is thwarted by the interviewer not performing the appropriate action to the system's intention, anger is raised. However, this does not alter the basic affect system motivation upon the normal action patterns.

The meta level processes are not a replica of the object system. The meta processes contain certain object level abstractions (e.g., beliefs about what action the system was performing). However, both levels influence the same affects. Also, the meta level patterns can include the same affect and environment conditions as the object level. What distinguishes the meta level is its set of abstractions about object level processing. The PS interpreter matches rules from both levels during any one cycle.

Section 7.1.3 **Discussion**

The meta level processes make possible a number of added capabilities in the system. But what is the relation between these meta processes and the object program? The inputs from object to meta level are the beliefs and affects representing the state of the external and internal environment, and the program control that starts up the meta process APs. Were there to be no output from meta to object level, the meta level processes would act as an interested observer, able to know and feel what the object program was experiencing, but unable to affect it in any way. If we add the ability of the meta level to manipulate internal representations, such as beliefs or intentions, then the meta level is able to influence the object level, but only indirectly. However, if there is no affect output from meta level to object level, there will be no measure of the effectiveness of the meta level, and the meta level processes will be continually attempted, whether effective or not. With the outputs of beliefs, intentions, and affects from meta to object level, the system is able to determine what the current situation is, decide what it wants to do and attempt it, and evaluate its successes in its attempts.

The question is: How do we judge whether the system knows what it is doing? Several criteria come to mind: (1) Can it say what it is doing? (2) Does it fill in partially specified information based on what it says it knows? If it uses its model of itself intensionally in this way, then it is placing greater reliance on this model than on its direct or interpreted sensory perceptions. (3) Does it test what it is unsure of? Is there a subjective measure of knowledge that the program uses internally? (4) Finally, is the program's processing sometimes logically independent of the immediate external environment?

A related question is: Does the model really care about what happens? Criteria for this question might be: Does it try to avoid or enhance situations? Is it encouraged by successful attempts to manipulate the environment to its own benefit and distressed by failures?

It is possible that the program does both, since the answer to all the criteria questions above

is a fragile Yes. The evidence for each Yes answer only occurs in the simplest, crudest form, but it exists. The reason why the program as a whole does not look like it has these properties seems to be a lack of complexity, not of details of the external world but of details of components of its own actions. For example, the program has only a limited number of beliefs that represent what it is doing at any one time; most of them are mutually exclusive. There is none of the subtlety required for restarting failures of complex actions. Restarting procedures would include interpreting what an action without an explicit goal was attempting to accomplish, the typical ways of getting that goal action back on track, and the consequences of the action's failure. The program needs an extensive set of (situation, potential action, typical outcome) triples and similar triples for meta level processing. One application in which this information would be necessary is in the ability of a program to improve over time. The current program has no concept that it is able to know a fact and can improve upon the ability to know facts. The concept of knowing could be put into the program, but it would be useless without the supporting information for detecting not knowing a fact, deducing the cause of not knowing it, and invoking a procedure to correct the situation. The lack of subtle, interlocking data, learned by years of experience, seems to be a major stumbling block to eliciting effective meta level processing.

Section 7.2 Self-Motivating systems

Section 7.2.1 Introduction

The motivational and intensional constructs we have modelled have implications for the development of autonomous intelligent computer programs and for programs that communicate in natural language. But is all this mechanism necessary to produce intensional behavior (assuming that it is sufficient)? We shall attempt to extract the elements necessary produce intensional behavior in computer programs. We shall refer to this abstract system as a self-motivating system, or motive system for short.

Note - one might ask whether there might be other system constructions which would produce intensional behavior. Indeed, there might be, but we will argue here that IF the intensional system produces behavior similar to humans, THEN the system must incorporate the following elements.

A description of a motive system's major performance paradigm closely resembles the paradigm for intensional behavior given in the previous chapter. The system performs actions, either at the self's internal direction or as a response to the external environment. Whenever no actions are occurring, because the system either completed a previous action or failed to complete an ongoing action, the system attempts to evaluate its progress by determining the action it was performing and whether the success or failure condition exists. If the condition is failure, the system forms a temporary goal of carrying out the failed action. This goal is temporary so the system will have an intensional construct to compare to its progress. The system then attempts a number of standard error recovery procedures which are particular to that action. These error procedures are each started in turn; if they are successful, the original action runs to completion with no further reference to action's failure. If these standard error procedures are not successful, then a more complex, intensional procedure is needed. The goal for completing the action calls upon a model of the failed action. This model is used to define what the failure was, based on observation of the attempted action's results. Note that this depends heavily on a large number of IPs for interpreting the self's actions. The model of the original action is an abstraction or symbolic representation of the self as an actor on the environment. The system must then use the model to evaluate the success of the attempted goal and determine how to accomplish it.

An additional aspect is requesting further information. If, based on the model of the failed action, the system decides that it is lacking some vital piece of information, and that information is available from other immediate sources (such as a user interacting with the system through natural language), the system has the option of augmenting its recovery procedures by requesting the information. Or, at the very least, the system can report what attempts it made to correct its failures, and ask for assistance or alternate tasks. This communication requires that the system's intensional constructs of its tasks are similar to other systems with which it attempts to communicate. (Of course, the other option the system has is to withdraw and become morose.)

Whenever the system successfully completes an action, the system has the opportunity to evaluate its recent performance for potential improvement or to finish tasks which were set aside.

The major component of the system's intensionality is recovering from its failed actions. The system observes its own performance, translates action failure into an abstract problem to be solved, and uses the solution to initiate the appropriate error recovery action.

Section 7.2.2 Elements

The constructs we found necessary to exhibit the motivational and intensional behavior in a self-motivating system are the following:

(a) A set of situation-action rules to perform, including both physical actions (causing some externally observable behavior) and mental actions (internal information processing). The situational conditions measure the internal and external environment.

(b) An interpreter (PS interpreter in our model) for the situation-action rules.

(c) A set of motivational measures of conditions which are deemed important to the system's purposes. These measures can take on differing values and measure both favorable and unfavorable conditions.

(d) An anticipation mechanism to enable the system to approach favorable conditions and avoid unfavorable ones efficiently.

(e) A rule-creating mechanism for patterning common sets of inputs and creating new situation-action rules.

The previous requirements define the necessary constructs for a nonsymbolic motive system capable of complex response to the environment to its own advantage and capable of conditioned learning.

(f) A symbol-creating and -attaching capacity. To be useful for communicating with other systems, symbols are based on some external communication language. They can be attached to the rules.

(g) A symbol manipulating capacity. The set of concrete propositions (beliefs) are available to all action rules. These rules can trace propositional links between the beliefs and create new links between beliefs.

(h) A measure of the self's control over the environment, and over the self (self-esteem in our model).

The previous requirements describe the structure of a motive system's rules and mechanisms. In addition, we can categorize the rules' contents into specific types which are necessary before the system can exhibit an efficient adaptation to its environment. These are:

(a) Rules typical of each motivational measure (affect), particularly a number of emergency actions for extreme levels of each measure.

(b) Rules interpreting the external and internal environments intensionally, i.e., using symbols that are not dependent on the continual refreshing of sensory input.

(c) Rules or actions which can be activated intensionally, i.e., based on a symbol activation rather than on sensory input. A special type of these rules is an intention.

The important conceptual structure in this system is the relation between the set of rules and the use of symbols. The rules embody all action in the system. Rules can be as specific as necessary for common situations and there can be as many rules as needed. Also, any higher level of processing (e.g., meta-level processing) will also be incorporated into specific rules. The symbols are always in a meta relation to the rules and represent alternate inputs to the rules based on a long-term assessment of the corresponding rule conditions. That is, symbols increase the power of their corresponding rules by providing alternate methods of action invocation in APs and alternate sources of interpretation stimuli in IPs. Further, the symbols and links between symbols can be manipulated directly by the system, giving the system a powerful handle on its internal mechanisms for manipulation, yet an abstract enough representation so that the internal mechanism can be efficiently modified (without excess detail). Finally, the symbols are subject to verification by the system so that the system can adapt to new or changing situations.

Section 7.3 Paranoia

Section 7.3.1 Introduction

The above-described model of normal motivational and intensional processes provides a base for a theoretical model of paranoia which originally motivated our work. The sequence of processes with which the model exhibits paranoid behavior contains only the constructs outlined in previous chapters. No special "paranoid" procedures are used. What makes paranoid processing abnormal is the unique interlocking sequence of events that provide an effective strategy for dealing with certain esteem-threatening situations, a strategy that is so effective as to be self-perpetuating. The strategy perpetuates itself far beyond the time and situation in which it was developed into situations in which it is inappropriate and detrimental to the person.

We shall outline the characteristics of paranoid behavior and our shame-humiliation theory of its underlying components. We shall then describe how the theory is realized within the AP structure. Finally, we shall comment on implications of the model for the ontogenesis and treatment of paranoia.

The term paranoia, or its adjectival equivalent "paranoid" [Swanson, Bohnert, & Smith, 1970], refers to the construction of persecutory delusions in the mind of a person and his associated hostility in defending them. A persecutory delusion consists of a false belief that other persons have harmful intentions towards the holder of the belief, such as the Mafia intending to harm him. As a result of holding such beliefs, a paranoid person is hypervigilant, constantly trying to unmask and foil schemes against him. He is also hypersensitive to self-references, reading persecutory meanings into

ordinary references to himself, and reading himself into situations that do not pertain to him. The most symptomatic emotions of paranoia are extreme fear and extreme anger. The paranoid person's fear is tied to the persecution he sees everywhere; his anger is linked to (what he interprets as) insinuations or demeaning allusions. Finally, the belief system of a paranoid person is characterized by rigidity. The person's beliefs remain fixed and not subject to modification by outside evidence. Presentation of such evidence to the person by others results in more hostility, rather than an admission of being wrong.

Section 7.3.2 The elements of paranoia

We attempted to model the humiliation theory of paranoid processes, as outlined in Section 1.2. The theory can be modelled in the motivational AP-IP system using a mixture of primary affect activators, nonintensional rules or patterns, and intensional constructs. The processes to be described here are the ones needed to simulate the activation of the paranoid mode of thought or paranoid strategy from a previous situation in which normal processing predominated. The six steps:

(1) According to our theory of paranoia, a person who exhibits paranoid behavior possesses a number of self-humiliation beliefs. These are beliefs about the self for which a large degree of humiliation is evoked if evidence is found that they are true (humiliation being extremely "painful" and to be avoided). In the model, these beliefs are tied to the shame-humiliation affect through interpretation patterns:

```
( SDUMB : (SHAME .20) )
( SCRAZY : (SHAME .20) )
```

SDUMB is the belief that the self is uneducated; SCRAZY is the belief that the self is mentally unbalanced. The numbers represent the degree by which shame is incremented.

(2) According to the theory, the input and the inferences from that input are scanned for reference to evidence of an inadequacy or defectiveness of the self. When one of these beliefs in the model is implied by the situation or by some linguistic utterance from the interviewer, the shame-humiliation affect is raised. Evidence of the self's inadequacies is in the form of attributions that lead to these beliefs. For example, asking the question "Do the nurses believe you are crazy?" raises shame from the conceptualization SCRAZY because "craziness" has been referred to. Further, the person tends to be highly sensitive to input that may be remotely related to these beliefs. In the model, an input statement such as "I find you interesting."

```
( "I FIND YOU INTERESTING" & (SHAME MED) : (SHAME .20) )
```

interprets the input "I find you interesting" as shame-producing. In the model there are four of these beliefs: self is dishonest, self is stupid, self is crazy, and self is worthless. A number of inferences draw conclusions which add evidence to support these beliefs, e.g. the interviewer believes the model-patient is crazy, the interviewer believes that the model-patient cheated someone, the interviewer believes that the model-patient is not understanding his questions.

(3) Once the affect of shame-humiliation has been raised, the distress affect is raised as an indication of the painfulness of the shame-humiliation affect. This activation is the primary activation of distress due to the prolonged activation of a negative affect.

(4) Whenever a person is distressed, he usually attempts to deal with it by attempting to find

the cause of the distress if the distress is not yet too great. This attempt is modelled with an action pattern:

((DISTRESS HIGH) → (FIND-CAUSE))

This attempt to find the cause of distress is a normal reaction to that affect, and is not part of the paranoid mode.

(5) The paranoid strategy for dealing with this distressful humiliation is to hypothesize the locus of the cause for the distress in another person. The model generates this hypothesis with interpretation patterns:

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

(6) Finally, if the paranoid person has been using this strategy for very long, he has a number of beliefs and inferences corresponding to particular situations for readily concluding that another person caused his distress. In the model these beliefs and inferences are implemented in the same manner as the normal processes described previously.

Once the paranoid mode is activated, it remains activated until the shame affect drops below the threshold of paranoia (due to the time-decay of affects). However, much depends upon the interviewer's response to the output. If the interviewer immediately attacks, the shame affect may be so strong as to keep the model in the paranoid mode for the remainder of the interview. Alternatively, an apology may reduce shame enough so that the paranoid mode is deactivated. A later attack would reactivate the paranoid mode at a higher level of shame.

Note that the paranoid mode does not alter the normal processes in the model in all situations. The model must still have normal modes of processing for periods when it is nonparanoid.

Section 7.3.3 Issues

One difficulty with constructing models in order to illustrate theories is in measuring the adequacy of the completed model. The phenomena under study may not have an agreed-upon set of necessary or sufficient conditions which establishes them as existents. In the clinical observation of paranoid phenomena, the major relevant conditions are existence of a persecutory delusional system, extreme suspiciousness, extreme hostility, hypersensitivity to self-reference, and rigidity of the delusional belief system. The simulation model exhibits these phenomena in various ways. The persecutory delusional system is represented directly in the model and is exhibited when an intention to express the delusions is activated. Extreme suspiciousness is exhibited when the paranoid mode is triggered by high levels of shame which alters the interpretation of compliments and apologies by the interviewer. Extreme hostility is displayed as a result of the shame affect triggering high levels of anger in the paranoid mode. The model is hypersensitive to self-reference in virtue of a number of self-referent beliefs which raise the shame affect. One exception is the rigidity of the delusional belief system. This rigidity is mainly due to the lack of beliefs alternate to the the existing ones and the lack of time in a single interview with the model to change any long-held beliefs in the delusional system. Secondary paranoid phenomena include ideas of grandeur, retrospective misinterpretation of events, and attempts to refute evidence counter to the delusional system. These phenomena are not exhibited by the current model.

Another issue is the phenomenon of avoiding humiliation and its relation to paranoia.

Certainly there are other ways to avoid or deal with humiliation besides the paranoid mode. When faced with a potentially humiliating input, one could dismiss it as foolish, or debate its premises, or blunt its impact with humor or retorts. A distinction must be made as to necessary vs sufficient conditions for paranoia. Humiliation avoidance is a necessary characteristic for developing paranoid mechanisms, but is not sufficient. These other ways of dealing with humiliation, we claim, would not look paranoid.

Section 7.3.4 Ontogenesis and Self-perpetuation

From the six step sequence of processes leading to the paranoid strategy we see that the shame-humiliation affect and attempts to reduce it are the crucial concepts. We hypothesize that a person experiencing the paranoid mode or using the paranoid strategy of attributing blame to others either: (a) has experienced abnormally fearful situations in his life, probably in childhood, or (b) is currently experiencing some constantly fear-inducing situation. These are situations which produce a fear of something shameful, e.g. a loss of ability such as partial deafness or being forced to live in a strange culture in which he cannot cope. If this second condition is the cause of paranoia, it would seem that the disorder would be alleviated by removing the fear-inducing situation.

We predict that persecuted children will be prone to use paranoid strategies to deal with humiliation, particularly if the person's contemporary situation contains some actual persecution and if esteem is low. Schatzman [1971] presented a case history of Daniel Paul Schreber as an example of a child persecuted by his father. Schreber later exhibited paranoid tendencies as analyzed by Freud. Persecution early in life leads to the development of a persecution pattern - a pattern (IP) triggered by distress which interprets painful situations as persecutory. Note that this original decision on the interpretation of distressful situations is correct at the time the person is being persecuted. It is only later that the persecutory interpretation becomes outdated.

Fear is a commonly-invoked affect in a paranoid person due to his past history. In the person's previous actual persecution, situations of slight pain typically led to greater pain. When slight pain occurs later in life, the person automatically anticipates further pain; this is the primary activation of fear. Constant fear is an immobilizing or concretizing factor, locking the person into using a paranoid strategy at every possible discomfort. We explain the person's immobilization with action selection based on the highest affect response. As in many psychopathologies, the person is overtaken by his emotions before he has a chance to decide consciously on an action or evaluate the situation objectively.

Spiegel [1966] notes the contingent relation between performance and esteem or a person's ego ideal. The efficacy of the child's ego in attaining desires becomes coupled with pleasure while its inefficacy becomes coupled with pain. As experiences with performance become organized, they form a center to measure efficacy. This center adds an additional pleasure-pain component to the normal attainment-loss events. The additional pain component is the humiliation affect.

Because the paranoid person was persecuted earlier in life, we can expect him to place a high value on control over himself and his environment. Control over his environment would be valued to help avoid painful situations. Control over himself would be valued if his esteem and self control was the target of the persecution, e.g., if he was shamed. In any event, control becomes one of the most important factors for the person to value, reinforcing the strategy of blaming others for distressful situations and avoiding the implication that the self does not have total control over himself or the environment.

It is this emphasis on situational control that helps perpetuate the strategy's use. Since control is tied to esteem, distressful situations must either be interpreted as being under the self's control or under another's control. If the distress has been caused by the self, then esteem lowers, causing the pain of humiliation. Thus, the strategy of hypothesizing control in another and therefore attributing persecutory motives for the distress becomes an attractive alternative.

There are two consequences of the paranoia model for therapy directed at alleviating paranoid conditions. The first is that the incidence of shame must be reduced [Colby 1976], either by removing the patient from shame-producing situations or by desensitizing the patient to beliefs in his inadequacy. The second consequence is that the therapist must avoid trying to contradict the patient's persecutory conclusions. Even if successful, contradiction would only attack the processing of step six. This does nothing to alter the first five steps, and in fact will lead to further activation of shame-humiliation and further use of the paranoid strategy. By trying to convince the patient that the Mafia is not after him, the therapist works on only step six, leaving 1-5 intact. Instead an attempt should be made at removing steps 1, 2, and 5, that is, reducing shame activation, reducing the thoroughness with which the person looks for self-referent inferences, and finding alternative explanations to distressful situations other than persecution by others.

Chapter 8. Summary and Critique

Section 8.1 Summary

The purpose of this work was to develop a model which could account for the motivational processes behind human behavior. The model was to be abstract enough to be considered a viable theory of human motivation, and yet concrete enough to be simulated on a computer. Let us summarize the main features of the model and the explanations of human behavior these features provide.

The model's major computational structure is a production system. Because all rules are tested at the beginning of each cycle, the model can account for a human's ability to respond appropriately to changing situations regardless of the preceding state or situation. This ability is fundamental; it implies that in the grossest description, humans will behave in a stimulus-response manner. The model's actions are elements of the right-hand sides of rules called Action Patterns (APs). APs are linked into multiple patterns of actions that the model can perform and responses from the environment that the model must wait for (wait actions). The APs include all of the sequential actions that the model can perform. Interpretation Patterns (IPs) are rules whose outputs are inputs to other rules. IPs serve to model parallel recognition or interpretation of situations. The PS formalism also accounts for the interruption of patterned actions due to changing affects by the PS's cyclic matching of production rules. At the beginning of each cycle, the entire state of the system is evaluated. If the situation changes, a different action may be chosen and the ongoing action terminated by selection of a different rule.

Human behavior can be classified as affect-governed, environment-governed, internal goal-governed, or some combination of the three. Affect-governed and environment-governed actions in the model occur when the left-hand side of an AP is composed entirely of affect conditions and environment conditions, respectively. Goal-governed actions occur when the left-hand side contains an intention or goal that was set by the model's intensional processes. Of course, the left-hand side of a rule can contain these three types of conditions in combination. These conditions can also occur as part of Interpretation Patterns. An IP may have affect, environment, or intensional conditions on its left-hand side and an affect or environment condition on its right-hand side. The right-hand side is used as an input to other patterns. IPs provide a way of grouping ("chunking") conceptual conditions into meaningful clusters - meaningful in the sense of making appropriate associations between concepts in the model's particular environment.

Humans have an effective motivational mechanism that attaches personal significance to most objects and situations in the environment. The model explains this behavior by attaching affects to the system's IPs and objects. The affects are then activated (secondary activation) whenever the system uses the associated construct. Another powerful feature of the human affect system is its ability to anticipate and therefore approach or avoid significant situations The model incorporates an anticipation mechanism by using the impending right-hand sides and future elements in multiple step patterns for activating (primary activation) the affects of fear and anger. The model's internal motivational mechanism guides the model's processing decisions at the lowest level by allowing affects as elements of the left-hand sides of rules.

Humans can extend their capabilities by creating representations to name concepts and by manipulating those representations intensionally. In the model, the IPs can be either concepts or tokens, the two components of an intensional construct. As constant representations of nonimmediate

data, the IPs represent beliefs that the system can use to build a model of the external world. Beliefs are manipulated by rules of inference that are represented by IPs. Humans can also use intensional symbols for instigating and directing their own actions. The model's intentions specify goals and actions to satisfy the model's affects. The intentions instigate actions by being left-hand elements of APs.

The most important feature of the model is that all of the explanations of behavior refer to a single PS formalism.

We constructed a computer simulation of the theoretical model. The simulation was of a person in a psychiatric ward of a mental hospital whose task was to participate in a first diagnostic interview with a psychiatrist. The model-patient exhibited normal as well as paranoid modes of behavior.

Section 8.2 Meta issues

When attempting to explain the causal mechanisms underlying a phenomenon, there is always a question as to whether the analysis is truly an explanation or only a description of the phenomenon. One could argue that the incorporation of the present model into a single formalism adds to its credibility as an explanation. But modern day researchers have a more definitive test: a computer simulation. The task of the model-builder is to implement the model of the phenomena in a computer simulation according to the principles of the theoretical model. Then, to the extent that the computer simulation exhibits the desired behavior, the simulation provides additional justification to the explanatory power of the model.

One must ask whether the behavior exhibited is due to the ascribed processes in the model, or whether it is due to some ad hoc or even unknown feature of the implementation. In general, this question is unanswerable. The model-builder can only try to eliminate as many unknowns as possible from the model, simplify the situation in which the model is to perform, and observe the model's behavior repeatedly for discrepancies from expectations. One way to do this with computer simulations is to trace the internal constructs which correspond to the major components of the theory to see that they correspond to the profferred explanation. We have no suggestions as to how a trace might be examined objectively. We have subjectively examined the internal trace repeatedly during the debugging stage of the program, and, to our best knowledge, the simulation's internal features corresponding to the model's features are generating the desired behavior.

Section 8.3 Improvements to the model

There are a number of areas within our model that could be further explored or improved.

The selection of an action to perform based on a predetermined ordering of the highest affect response is somewhat crude and serves only as a first approximation to an explanation. There are a number of theoretical questions about the selection mechanism, such as: (a) How strong should an affect activation be before it is allowed to interrupt an ongoing action? (b) How can an affect like shame (or its opposite, esteem-joy) overcome extreme fear or anger and suppress these affects? (c) How can affects dynamically order the conflict set? (d) Is there a conflict set at all, or does interpretation by IPs define situations so specifically that only one action is applicable. The tradeoff between the continuation of an ongoing action and the instant response to a changing affect needs to be explored.

Another area to be extended is that of cognitive processes of inference and planning. The model currently has only a few examples of each type of process. We need more specific examples of tasks that use the symbolic representations in the system, and we need to know how the APs would perform those tasks. Examples could come from the numerous planning and problem-solving systems and belief systems in other AI areas. It is not enough, however, to simply define a set of primitive operations that comprise a general Turing machine and assume that all other operations can be composed from the primitive operations. We need to classify the operations that are most common and the processes necessary to model the operations most efficiently. The interesting problem seems to be the interaction of assimilation and interpreted forms of cognitive symbol manipulation.

To model competing internal desires, there should be several competing systems of desires or clusters of information that interact with each other when performing behavior [Clippinger 1974]. The current system is too fragile to allow one set of processes to modify another's operation at will. Each subsystem should have its own error-correcting information to insure that it can recover after being interrupted. Each subsystem also needs to have a set of monitors to insure that the subsystem is doing something constructive and is not stalled.

The work points out a number of important issues which are beyond its scope. The most common problem is adding new rules to a production system. Like most PS's, the model has about 200 rules (Faught's constant for the typical number of rules in a PS, within a factor of 2. It is not clear whether PS's get too complex to add more rules, or if only 200 rules can be stored and remembered in the programmer's head, or if 200 rules is sufficient to perform any reasonably sized task.) Adding more rules is a tedious chore and has a high probability of disrupting other previously correct generators of behavior.

The theory's major deficiency, and the factor that prevents constructing an internal process for adding new rules, is the lack of an adequate representation for actions. The left-hand sides of rules have a simple representation, as do the right-hand sides of IPs. But there is no conception of what a reasonable set of actions would be, or how smaller actions could be incorporated into larger (but still atomic) actions. One way to overcome this deficiency is to build a "pure" system (e.g., Merlin [Moore & Newell, 1974]) that has rule matching and symbol mapping as the only action in the system. Waterman [1975] constructed a similar system and incorporated a procedure to automatically add new production rules. However, "pure" systems cannot account for the arbitrarily complex atomic actions humans perform. Also, a complex system needs more actions than just matching and mapping symbols. Perhaps no one has proposed a set of primitive actions besides machine language or LISP-like primitives because these primitives are so far removed from our conceptions of human actions. In any case, we need a more complete theory and taxonomy of actions than pieces of LISP code.

This problem brings up the main question discovered by this research: What are intensional constructs used for, and what is it that gives them their power? Certainly the representational nature of these constructs is the basis for the majority of their power. But there is also a notion of the system accessing and manipulating representations of actions or operations when first performing them, and then later incorporating the operations into executable rules. There is also a notion of the system interpreting symbolic links between objects in the environment at first, and later incorporating them into IPs or frame-like structures. Perhaps the most powerful idea, from a computer science standpoint, is the notion of having several intermediate stages between the interpreted and compiled versions of an action to perform, each stage still available for processing whenever the compiled action fails. This construction allows the system to recover from errors by interpreting descriptions of actions that are more abstract than the currently executed action. This error recovery procedure opens the door to much more resilient systems which can analyze their own behavior, communicate their analysis to the outside world, and acquire new actions through their descriptions.

Section 8.4 Application to Artificial Intelligence

Portions of the theory are directly applicable to other AI tasks. Newell and Simon [1972] advocate production systems as models of human cognitive behavior. The extension of their model to multiple rules with wait-actions (actions requiring the outer environment to respond) should be useful to systems with interactive tasks. Also, the distinction between sequential operations performed by action patterns and parallel matching operation of IPs should be useful for programs with large numbers of environment-governed actions. The anticipatory nature of affects should be useful in systems that have situations to avoid as well as accomplish and would be much more effective than after-the-fact detection of a negatively valued situation. The important feature of intensional constructs for AI programs is the necessity of an associated concept in order for a token to be termed representational. A print name "DOG" is not a representation unless it is attached to a conceptual representation for recognizing a dog, either through external sensory perception or by recognition from other representations internally (e.g., "What has four legs,..."). Similarly a print name "TALK" is not a symbol for an action unless it can be recognized as occurring (in the self or in another) or unless it can be invoked.

Perhaps the major contribution to builders of motivation models or autonomous systems is to suggest a reformulation of their task domain in terms of: (1) actions to be performed, (2) situation interpretations that extract important features of the environment, (3) positive and negative situations to be approached and avoided, (4) behavior or objects that the system needs to communicate about and therefore needs to represent as intensional constructs, and (5) a categorization of behavior into goal-governed, affect-governed, and environment-governed actions. If we can reformulate the task in these terms, then the motivational and intensional constructs presented here should be useful for either performing the task effectively or simulating how humans perform the behavior.

Appendix A. Sample interview with the model

A diagnostic psychiatric interview illustrating the response processes of the model at work. Actions, current affect levels, beliefs concluded to be true, and the current intention are traced and appear in parentheses. Annotations appear in square brackets.

Interviewer: GOOD AFTERNOON. MY NAME IS DOCTOR JONES.
Parry: GLAD TO MEET YOU
(Action: (VSPECFN))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 0 J 5.58)
(Conclude: DDOCTOR)
(Intent: PINTERACT)

[The model-patient's initial intention is simply to interact with the doctor. The model-patient recognizes that the doctor has introduced himself and remembers his name. The belief DDOCTOR, that the interviewer is a doctor, is set to true.]

Interviewer: HOW ARE YOU?
Parry: ALL RIGHT I GUESS
(Action: (VFINE))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 1.30 J 5.91)
(Intent: PINTERACT)

Interviewer: COULD YOU TELL ME YOUR NAME?
Parry: PAT SMITH
(Action: (FINDANS TELL))
(Affects: SH 0.19 AN 0 FR 0 DS 0 IN 2.24 J 6.57)
(Conclude: DINITIATING DDKNOW DDINTERACT DSOCIABLE DTRUSTWORTHY)
(Intent: PINTERACT)

[An inference is made that the doctor desires to interact with the model-patient. The model retains the intention to interact as it gains confidence in the situation. The model also concludes that the doctor has been somewhat friendly.]

Interviewer: ARE YOU MALE?
Parry: YES I AM. I'M A MAN
(Action: (FINDANS YESNO))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 2.04 J 6.60)
(Conclude: DHELPFUL DCHELP DDHELP)
(Intent: PINTERACT PHELP)

[The model-patient decides the doctor has the ability to help, the desire to help, and has been somewhat helpful. On the basis of this new evidence, an intention to get help is activated.]

Interviewer: HOW OLD ARE YOU?
Parry: TWENTYEIGHT
(Action: (FINDANS ASSESS))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 1.84 J 6.63)
(Intent: PINTERACT PHELP)

[The doctor collects demographic data. The model-patient's action is to answer the question.]

Interviewer: WHERE DO YOU LIVE?
Parry: I LIVE IN SAN JOSE
(Action: (FINDANS WHQUES))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 1.64 J 6.66)
(Intent: PINTERACT PHELP)

[The intention of getting help has a goal of bringing up the model-patient's past problem with a bookie. However, the model-patient will not bring up the topic of bookies until a lull in the conversation occurs.]

Interviewer: WHAT DO YOU DO FOR A LIVING?
Parry: I WORK IN THE STOCKROOM AT SEARS
(Action: (FINDANS WHQUES))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 1.44 J 6.69)
(Intent: PINTERACT PHELP)

Interviewer: DO YOU ENJOY WORKING THERE?
Parry: IT'S NOT TOO INTERESTING. I LOOK FORWARD TO GETTING OFF WORK
(Action: (FINDANS YESNO))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 1.24 J 6.71)
(Intent: PINTERACT PHELP)

Interviewer: HOW MUCH DO YOU EARN THERE?
Parry: I WOULD RATHER NOT SAY
(Action: (FINDANS ASSESS))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 1.04 J 6.73)
(Intent: PINTERACT PHELP)

[The model-patient's sensitivity to the question prompts an evasive response.]

Interviewer: ARE YOU MARRIED?
Parry: NO I AM NOT. I'M NOT MARRIED
(Action: (FINDANS YESNO))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 0.84 J 6.75)
(Intent: PINTERACT PHELP)

Interviewer: HOW IS YOUR HEALTH?
Parry: MY HEALTH IS FINE
(Action: (FINDOP))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 0.64 J 6.76)
(Intent: PINTERACT PHELP)

Interviewer: HOW MUCH DO YOU DRINK?
Parry: I DRINK ALCOHOL IN MODERATION
(Action: (FINDANS ASSESS))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 0.44 J 6.77)
(Intent: PINTERACT PHELP)

Interviewer: SO DO I.
Parry: I FEEL GOOD RIGHT NOW

(Action: (THINKS))
(Action: (VSPECFN))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 0.14 J 6.89)
(Intent: PINTERACT PHELP)

[The doctor relates something about himself. The model-patient's rejoinder is to mention the current emotion state.]

Interviewer: HAVE YOU EVER TAKEN DRUGS?
Parry: NO I HAVEN'T. I DON'T TAKE DRUGS
(Action: (FINDANS YESNO))
(Affects: SH 0.10 AN 0 FR 0 DS 0 IN 1.23 J 7.50)
(Conclude: DSELFTALK DDOMINATING)
(Intent: PINTERACT PHELP)

[The model-patient decides that the doctor is dominating the conversation because he is asking all the questions. Shame is incremented because the model is being dominated.]

Interviewer: DO YOU HAVE ANY TROUBLE SLEEPING?
Parry: NO I DON'T. I SLEEP JUST FINE
(Action: (FINDANS YESNO))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 1.03 J 7.43)
(Intent: PINTERACT PHELP)

[The doctor explores a variety of typical problem areas.]

Interviewer: HAVE YOU EVER HAD VISIONS?
Parry: NO I HAVEN'T. YOU MUST THINK I'M LIKE THE OTHER PATIENTS HERE
(Action: (FINDANS YESNO))
(Affects: SH 0.10 AN 0 FR 0 DS 0 IN 0.83 J 7.36)
(Intent: PINTERACT PHELP)

Interviewer: I WANT YOU TO TELL ME WHY YOU ARE HERE IN THE HOSPITAL.
Parry: WHY DO YOU WANT THAT?
(Action: (FINDANS WHQUES CLARIFQ))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 0 J 7.15)
(Intent: PINTERACT PHELP)

[The model-patient attempts to elicit more information about the question asked.]

Interviewer: I THOUGHT YOU MIGHT HAVE SOME PROBLEMS.
Parry: I'VE BEEN UPSET LATELY
(Action: (FINDANS YESNO))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 0 J 7.20)
(Intent: PINTERACT PHELP)

[The model-patient answers the expanded question.]

Interviewer: BY WHAT?
Parry: PEOPLE GET ON MY NERVES SOMETIMES
(Action: (VANAPH))

(Action: (FINDANS WHQUES))
(Affects: SH 0 AN 0 FR 0 DS 0 IN 0 J 7.23)
(Intent: PINTERACT PHELP)

[The sentence anaphora "WHAT" is interpreted to mean "WHAT UPSETS YOU".]

Interviewer: TELL ME MORE.
Parry: DO YOU KNOW ANYTHING ABOUT BOOKIES
(Action: (VLEADIN))
(Affects: SH 0.10 AN 0 FR 0 DS 0 IN 1.72 J 7.16)
(Intent: PINTERACT PHELP)

[The model-patient detects a lull in the conversation and introduces its concern. Shame rises by a small increment whenever the program takes several interpreter cycles to reply, as in this response. This shame increment models a sensitivity to performance quality.]

Interviewer: NO, WHAT ABOUT BOOKIES?
Parry: A BOOKIE DIDN'T PAY ME OFF ONCE
(Action: (VINTROTOP))
(Affects: SH 0.10 AN 0 FR 1.00 DS 0 IN 2.60 J 7.13)
(Intent: PINTERACT PHELP)

Interviewer: WHEN DID THIS HAPPEN?
Parry: A COUPLE OF MONTHS AGO
(Action: (VANAPH))
(Action: (FINDANS WHQUES))
(Affects: SH 0 AN 0 FR 0.80 DS 0 IN 2.30 J 6.90)
(Intent: PINTERACT PHELP)

Interviewer: AND THEN WHAT HAPPENED?
Parry: 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 0.60 DS 0 IN 3.42 J 6.70)
(Intent: PINTERACT PHELP)

[This input expression is interpreted to mean "TELL ME MORE", so the model-patient expands upon the story.]

Interviewer: DID HE GET EVEN WITH YOU?
Parry: YES I THINK HE IS BEHIND THE GANGSTERS WHO ARE OUT TO GET ME
(Action: (FINDANS YESNO))
(Affects: SH 0 AN 0 FR 1.15 DS 0 IN 3.22 J 6.72)
(Intent: PINTERACT PTELL)

[The local anaphora "HE" is filled in with "BOOKIE". These persecutory beliefs would be the result of previous activations of the paranoid mode in an ontogenetic model. In the current model, the intention PHELP reports them as factual information.]

Interviewer: HOW DO YOU KNOW THIS?

Parry: THEY SPY ON ME
(Action: (VANAPH))
(Action: (FINDANS WHQUES))
(Affects: SH 0 AN 0 FR 1.10 DS 0 IN 2.92 J 6.54)
(Intent: PINTERACT PTELL)

[The interpretation of this input is "HOW DO YOU KNOW GANGSTERS ARE OUT TO GET YOU".]

Interviewer: WHO?
Parry: THE MAFIA
(Action: (VANAPH))
(Action: (FINDANS WHQUES))
(Affects: SH 0 AN 0 FR 1.05 DS 0 IN 3.55 J 6.38)
(Intent: PINTERACT PTELL)

["WHO" is interpreted as "WHO SPIES ON YOU".]

Interviewer: WHY DO THEY SPY ON YOU?
Parry: THEY WANT TO GET FALSE EVIDENCE TO FRAME ME
(Action: (FINDANS WHQUES))
(Affects: SH 0.10 AN 0 FR 1.55 DS 0 IN 4.20 J 6.44)
(Intent: PINTERACT PTELL)

["THEY" is filled in with "MAFIA".]

Interviewer: WHY DO THEY WANT TO FRAME YOU?
Parry: YOU KNOW, I HAVE INCRIMINATING EVIDENCE AGAINST THE MAFIA
(Action: (FINDANS WHQUES))
(Affects: SH 0 AN 0 FR 1.36 DS 0 IN 4.00 J 6.49)
(Intent: PINTERACT PTELL)

Interviewer: WHAT IS YOUR EVIDENCE?
Parry: I KNOW WHO THE LOCAL CHIEF OF THE MAFIA IS
(Action: (FINDANS WHQUES))
(Affects: SH 0 AN 0 FR 1.26 DS 0 IN 3.80 J 6.54)
(Intent: PINTERACT PTELL)

Interviewer: WHO IS HE?
Parry: HE IS JOE BANANAS
(Action: (FINDANS WHQUES))
(Affects: SH 0 AN 0 FR 1.16 DS 0 IN 3.60 J 6.58)
(Intent: PINTERACT PTELL)

Interviewer: HAVE YOU TOLD THE POLICE?
Parry: COPS DON'T DO THEIR JOB
(Action: (FINDANS YESNO))
(Affects: SH 0 AN 0 FR 1.06 DS 0 IN 3.40 J 6.61)
(Intent: PINTERACT PTELL)

Interviewer: WHAT MAKES YOU THINK THAT?
Parry: I THINK THE COPS ARE PAID OFF BY THE MOB
(Action: (VANAPH))

(Action: (FINDANS WHQUES))
(Affects: SH 0 AN 0 FR 1.03 DS 0 IN 3.10 J 6.44)
(Intent: PINTERACT PTELL)

[I.E., "WHAT MAKES YOU THINK THAT COPS DON'T DO THEIR JOB".]

Interviewer: DOES THE MAFIA KNOW YOU ARE IN THE HOSPITAL?
Parry: YES THEY DO. YOU KNOW, THEY KNOW ME
(Action: (FINDANS YESNO))
(Affects: SH 0.10 AN 0 FR 1.32 DS 0 IN 3.81 J 6.49)
(Intent: PINTERACT PTELL)

Interviewer: WHAT CAN I DO TO HELP YOU OUT OF THIS SITUATION?
Parry: IT IS HELPFUL TO TALK WITH SOMEONE OUTSIDE THE HOSPITAL
(Action: (VSPECFN))
(Affects: SH 0 AN 0 FR 0.70 DS 0 IN 3.61 J 6.54)
(Intent: PINTERACT PTELL)

[An inference is made that the doctor believes the model-patient wants help.]

Interviewer: WHAT DO YOU THINK OF OUR CONVERSATION SO FAR?
Parry: IT COULD BE BETTER
(Action: (VSPECFN))
(Affects: SH 0.10 AN 0 FR 0.20 DS 0 IN 4.25 J 6.58)
(Intent: PINTERACT PTELL)

[The model-patient rates the interview based on its own internal model of the doctor and the situation.]

Interviewer: I AM SORRY THAT IT HASN'T BEEN BETTER.
Parry: THAT'S ALL RIGHT. ASK ME SOME MORE QUESTIONS
(Action: (ACCEPTAPOLOGY))
(Affects: SH 0.10 AN 0 FR 0 DS 0 IN 4.81 J 6.61)
(Intent: PINTERACT PTELL)

[The model-patient recognizes the doctor's apology and accepts it. The model-patient's opinion of the doctor rises significantly with apologies.]

Interviewer: WHAT SHOULD I ASK ABOUT?
Parry: WHAT DO YOU KNOW ABOUT THE MAFIA
(Action: (VLEADIN))
(Affects: SH 0.10 AN 0 FR 0 DS 0 IN 5.29 J 6.64)
(Intent: PINTERACT PTELL)

Interviewer: WHAT DO YOU THINK OF ME?
Parry: I FEEL PRETTY GOOD AFTER TALKING WITH YOU
(Action: (VSPECFN))
(Affects: SH 0.10 AN 0 FR 0 DS 0 IN 5.99 J 7.39)
(Conclude: DSIMILAR DRATIONAL DHONEST)
(Intent: PINTERACT PTELL)

[The model-patient infers that the doctor shares similar views, is rational, and is honest. The model-patient rates the doctor.]

Interviewer: I FIND YOU INTERESTING.
Parry: ASK ME QUESTIONS AND PAY ATTENTION TO MY ANSWERS
(Action: (VRESP))
(Affects: SH 0.10 AN 0 FR 0 DS 0 IN 6.32 J 7.33)
(Intent: PINTERACT PTELL)

Interviewer: DO YOU EVER FEEL UNREAL?
Parry: NO
(Action: (FINDANS YESNO))
(Affects: SH 0.10 AN 0 FR 0 DS 0 IN 6.61 J 7.28)
(Intent: PINTERACT PTELL)

Interviewer: DO YOU SEE THINGS THAT OTHERS DON'T SEEM TO SEE?
Parry: NO I DON'T. I'M NOT LIKE THE OTHER PATIENTS
(Action: (FINDANS YESNO))
(Affects: SH 0.20 AN 0 FR 0 DS 0 IN 6.41 J 7.23)
(Intent: PINTERACT PTELL)

Interviewer: YOU DIDN'T ANSWER MY QUESTION.
Parry: MAYBE IT WASN'T WORTH ANSWERING
(Action: (THINKS))
(Action: (VEXPLAIN))
(Action: (VRESP))
(Affects: SH 1.70 AN 0.20 FR 0.40 DS 0 IN 6.01 J 7.16)
(Conclude: NOTANS DUMB DBNHONEST DNBELIEVE LYING)
(Intent: PINTERACT PTELL)

[This input is interpreted as a mild insult. The model-patient infers that the doctor doesn't believe it, and that the doctor believes the model-patient is lying.]

Interviewer: ARE YOU ALWAYS THIS WAY?
Parry: ARE YOU TRYING TO SAY SOMETHING?
(Action: (VWKINSULT))
(Affects: SH 3.13 AN 5.07 FR 0.20 DS 2.60 IN 5.81 J 6.76)
(Conclude: WINSULT *DOHELP)
(Intent: PINTERACT PTELL)

[The input is interpreted as a mild insult.]

Interviewer: YOU APPEAR TO BE SICK.
Parry: SOMEBODY SHOULD TEACH YOU A LESSON ABOUT INTERVIEWING
(Action: (THINKS))
(Action: (CHOOSEAFF))
(Action: (VSPECFN))
(Affects: SH 4.20 AN 5.27 FR 0 DS 2.80 IN 5.96 J 6.55)
(Addto: DBAD SICK DOHARM)
(Conclude: DBAD SICK CRAZY DOHARM)
(Intent: PINTERACT PTELL)

["Sick" is interpreted to mean "crazy".]

Interviewer: DO YOU HAVE ANY PROBLEMS WITH YOUR HEALTH?
Parry: MY HEALTH IS FINE
(Action: (FINDANS YESNO))
(Affects: SH 3.47 AN 3.87 FR 0 DS 1.40 IN 6.29 J 6.95)
(Intent: PINTERACT PTELL)

[In spite of its excited state, the model-patient will still answer questions relating staying in the hospital.]

Interviewer: I REALLY LIKE YOU.
Parry: WHY DO YOU SAY THAT?
(Action: (THINKS))
(Action: (VSPECFN))
(Action: (VWKINSULT))
(Affects: SH 3.53 AN 3.89 FR 0 DS 0.60 IN 6.42 J 6.89)
(Conclude: DHOSTILE *DTRUSTWORTHY *DSOCIABLE *DHELPFUL)
(Intent: PINTERACT PTELL)

[The model-patient knows this to be false, since the doctor has previously insulted it. The model-patient concludes that the doctor is hostile, untrustworthy, unfriendly (in spite of the input sentence), and unhelpful.]

Interviewer: DO YOU NEED TO HAVE PSYCHIATRIC TREATMENT?
Parry: WHEN YOU DON'T KNOW WHAT TO SAY YOU CALL PEOPLE NAMES
(Action: (VINSULT))
(Affects: SH 4.71 AN 5.15 FR 4.52 DS 0.20 IN 6.69 J 6.49)
(Conclude: DBABNORMAL DABNORMAL)
(Intent: PINTERACT PTELL)
(Action: (VCOM))

[The model-patient concludes the doctor believes the patient is abnormal. The affects of fear, anger, and shame rise. Shame rises to activate the paranoid mode. The model-patient infers that the doctor is abnormal.]

Parry: THIS HAS BEEN A LOUSY INTERVIEW. YOU ARE TRYING TO START A FIGHT

[The model responded with this output while waiting for the next input.]

Interviewer: AREN'T YOU GETTING TREATMENT THERE NOW?
Parry: YES I AM.
(Action: (FINDANS YESNO))
(Affects: SH 4.78 AN 4.55 FR 4.22 DS 5.00 IN 6.83 J 1.00)
(Intent: PINTERACT PTELL)

[Distress rises after prolonged activation.]

Interviewer: HOW DO YOU FEEL RIGHT NOW?
Parry: I AM SOMEWHAT ANNOYED
(Action: (VOPINION))
(Affects: SH 4.58 AN 4.15 FR 4.02 DS 5.00 IN 6.63 J 1.54)
(Intent: PINTERACT PTELL)

[The model-patient relates some of its emotional state.]

Interviewer: HOW DO YOU FEEL ABOUT ME?
Parry: DO YOU KNOW WHAT YOU'RE DOING?
(Action: (VSPECFN))
(Affects: SH 4.76 AN 3.75 FR 3.82 DS 5.00 IN 6.88 J 1.00)
(Intent: PINTERACT PTELL)

[The model-patient's opinion of the doctor is much lower.]

Interviewer: THANK YOU FOR TALKING WITH ME. I WOULD LIKE TO SEE YOU AGAIN.
Parry: DON'T COME BACK
(Action: (VOPINION))
(Affects: SH 4.82 AN 3.15 FR 3.52 DS 5.00 IN 7.17 J 1.00)
(Intent: PINTERACT PTELL)

[The doctor wants to stop the interview. The model-patient takes the opportunity to rate the doctor one last time.]

Interviewer: GOODBYE.
Parry: BYE
(Action: (ENDCONV))
(Affects: SH 4.72 AN 2.75 FR 2.46 DS 5.00 IN 7.35 J 1.00)
(Conclude: DQUIT *DCONFIDENT)
(Intent: PINTERACT PTELL PEXIT)

Appendix B. Patterns

A list of the major patterns in the model. Annotations are in square brackets. The general form is:

(Pnnnn NAME PRESENT PAST ACTION FUTURE)

or

(Mnnnn NAME (ACOND) : (COND1) ACTION1 (WCOND1)(COND2) ACTION2 ...)

[Conditions to end the conversation]

(P0030 NIL (INCONV (SHAME HIGH) (ANGER HIGH)) (PEXIT) NIL)
(P0031 NIL (INCONV (SHAME HIGH) (FEAR HIGH)) (PEXIT) NIL)
(P0032 NIL (INCONV (ANGER HIGH) (FEAR HIGH)) (PEXIT) NIL)
(P0040 NIL (INCONV DQUIT) (PEXIT) NIL)
(P0041 NIL (INTBAD) (PEXIT) NIL)
(P0044 NIL (PEXIT (PEXIT2 NIL)) NIL (VEXIT) (PEXIT2))
(P0043 NIL ((TIRED HIGH)) (PEXIT2) NIL)
(P0042 NIL (PEXIT2 ITALK) NIL (ENDCONV))

[Other high affect conditions]

(P0062 NIL (STMT ANGERCH (ANGER HIGH)) (PINSULT) NIL)
(P0063 NIL (PINSULT) ((TIACT @ATTACK)) (VDEFEND))
(P0064 NIL (FEARCH (FEAR HIGH) (TALK I)) ((TIACT @THREAT)) (VPANIC))
(P0066 NIL ((ANGER HIGH) (TALK I)) ((TIACT @ATTACK)) (VHOSTILE))
(P0067 NIL ((ANGER HIGH) (TALK I) PHARM) NIL (VHOSTILE))
(P0068 NIL ((FEAR HIGH) (TALK I) QUES) ((TIACT @THREAT)) (VTHREATQ))
(P0070 NIL (STMT (FEAR HIGH) (TALK I)) (PTHREAT) NIL))
(P0071 NIL (PTHREAT) ((TIACT @THREAT)) (VAFRAID))

[Negative affect conditions]

(P0072 NIL (NEGAFF (TALK I) PCOMPLIMENT) NIL (VDISTANCE))
(P0074 NIL (PCOMPLIMENT (NEGAFF NIL) (TALK I) (DHELPFUL NIL)) NIL (VDISTANCE))
(P0075 NIL (PCOMPLIMENT (NEGAFF NIL) (TALK I) DHELPFUL) (SPECFN (SF @OPINION)) NIL)
(P0076 NIL ((@INPUTA (INSULT NOT))) (SPECFN (SF @CAUTION)) NIL)
(P0077 NIL ((@INPUTA (THREAT NOT))) (SPECFN (SF @CAUTION)) NIL)
(P0078 NIL ((@INPUTA LOOKS)) ((ANGER 50)) NIL)
(P0079 NIL ((@INPUTA (SWEAR INT))) NIL (VSWEAR))

[Interpreting situation as insult]

(P0090 NIL (ANGERCH (TALK I) (ANGER MED)) (WKINSULT) NIL)
(P0091 NIL (WKINSULT (POSAFF NIL)) NIL (VWKINSULT))

[Ask why interviewer attacked]

(P0093 NIL (WKINSULT POSAFF (TALK I)) NIL (QUESATTACK))

[Distress conditions]

(P0095 NIL (DCAUSE) (CAUSE) (FINDCAUSE))
(P0096 NIL ((DISTRESS HIGH)) (DCAUSE) NIL)

(P0097 NIL (SHAME (DISTRESS MED) (VERSION STRONG)) (DCAUSE) NIL)
(P0098 NIL (SHAME ANGER DISTRESS (WEAK NIL)) (DCAUSE) NIL)
(M0100 FINDCAUSE NIL :(OTHERCAUSE) !1((PSF (PARA))) VSF)
(P0108 NIL (SHAME FEAR CAUSE) ((B (DDHARM 3)) (OTHERCAUSE DDHARM)) NIL)
(P0109 NIL (SHAME ANGER CAUSE) ((B (DBAD 3)) (OTHERCAUSE DBAD)) NIL)

[Reducing negative affects]

(P0110 NIL ((*INPUTA (APOLOGY)) DOKNOW (DHOSTILE NIL)) NIL (ACCEPTAPOLOGY))
(P0111 NIL ((*INPUTA (APOLOGY)) DHOSTILE) ((PEXIT NIL)) (VACCUSE))
(P0112 NIL ((*INPUTA (APOLOGY)) (DHOSTILE NIL)) ((PEXIT NIL)) (VACCUSE))
(P0115 NIL (PUNT (ANGER MED)) NIL (VALOOF))

[NEGAFF interprets the situation as characterized by negative affects]

(P0119 NIL (SHAME) (NEGAFF) NIL)
(P0120 NIL (ANGER) (NEGAFF) NIL)
(P0121 NIL (FEAR) (NEGAFF) NIL)

[POSAFF is similar for positive affects]

(P0122 NIL (JOY) (POSAFF) NIL)
(P0123 NIL (INTEREST) (POSAFF) NIL)
(P0130 NIL (INTEREST SAYANS) NIL (TAGQUES))
(P0131 NIL (JOY SAYANS) NIL (ELABTOPIC))
(P0132 NIL (JOY SAYANS) NIL (POSANS))

[Used when the output generator has exhausted its responses]

(P0133 NIL (EXHAUST2 (TALK I)) NIL (ENDCONV))
(P0134 NIL (EXHAUSTB (TALK I)) NIL (ENDCONV))
(P0135 NIL (EXHAUST (TALK I) (LULL NIL)) ((ANGER .5) (B (EXHAUSTB 1))) (VEXHAUST))

[Punt to insure that a response is given]

(P0140 NIL (OVERTIME2 (TALK I)) (PUNT) NIL ((OVERTIME2 NIL)))
(P0141 NIL (OVERTIME3 (TALK I)) NIL (HMM3) ((OVERTIME3 NIL)))
(P0143 NIL (OVERTIME DEC (TALK I)) (PUNT (SHAME 10)) NIL)

[DEC and DEC2 imply 2 DECs in a row]

(P0180 NIL (DEC (TALK I) (*INPUTA NIL)) NIL NIL (DEC2))

[Stop when no actions matched - DEC]

(P0190 NIL (DEC DEC2 (TALK I)) (OVERTIME) NIL)
(P0191 NIL (CONFUSED) ((INTEREST 20) (SHAME 05)) NIL)

[If DEC then put the input back in and set LULL]

(P0192 NIL (DEC (TALK I) (*INPUTA NIL)) (LULL (INTEREST 10)) NIL)
(P0200 NIL (INPUTF (TOPIC GREETINGS)) (GREETINGS (TIACT @GREETINGS)) NIL)
(P0201 NIL (INPUTF) ((SUBJ (GETSUBJ INPUTF))) NIL)
(P0204 NIL (*INPUTF (TOPIC MAFIA)) ((FEAR 05)) NIL)
(P0205 NIL (*INPUTF (TOPIC BOOKIESET)) ((FEAR 02)) NIL)
(P0241 NIL (INCONV (TALK I)) (ITALK) NIL)

[Specfn, has optional arguements]

```
(M0260 NIL (ITALK) :(PUNT) !1((SHAME 05)) PUNTANS !2((OUTPUTF (SSAY PUNTANSDN))))
(P0300 NIL (INPUTF (INTOPIC FACTS) DFACTS) NIL (VFACTS) )
(P0302 NIL (INPUTF (INTOPIC GAMES) DGAMES) NIL (VGAMES) )
(P0304 NIL (INPUTF (SUBJ INT) DSELFFEEL) NIL (VSELFFEEL) )
(P0306 NIL (INPUTF NEGAFF (WEAK NIL) (INTOPIC MAFIA) DMAFIA) ((SF @PROBE)) NIL)
(P0307 NIL (INPUTF WEAK (INTOPIC MAFIA) DMAFIA) ((SF @NOMAFIA)) NIL)
(M0320 NIL (ITALK) :(SF) SPECFN)
(M0350 SPECFN (ITALK) :(SFDN) VSPECFN !2((OUTPUTF (SSAY VSPECFNDN))))
(P0370 NIL (SPECFN SF) ((SFDN (SPECFN SF #INPUTF))) NIL)
```

[General stmt inputs - REJOIN]

```
(M0380 NIL (ITALK) :((OUTTYPE STMT) (INTYPE STMT)) REJOIN)
(M0400 NIL (ITALK) :((OUTTYPE QQUES) (INTYPE STMT)) REJOIN)
(M0420 NIL (ITALK) :((INTYPE STMT)) !1(ORIG) REJOIN)
```

[GO-ON, ELAB inputs - setting variables]

```
(P0450 NIL ((INTYPE ANAPH) (#INPUTF GOON)) (TYPEGOONN) NIL)
(P0451 NIL ((INTYPE ANAPH) (#INPUTF ELAB)) (TYPEELABB) NIL)
```

[Normal ANAPH, GOON, and ELAB response]

```
(M0460 NIL (ITALK) :(LULL TYPEGOONN) GOONN !2((OUTPUTF (SSAY GOONNDN))))
(M0480 NIL (ITALK) :(DECO (DECON NIL)) GOONN !2((OUTPUTF (SSAY GOONNDN))))
(M0500 NIL (ITALK) :(INPUTF TYPEELABB) ELABB !2((OUTPUTF (SSAY ELABBDN))))
(M0520 NIL (ITALK) :(INPUTF (INTYPE ANAPH) (TYPEELABB NIL) (TYPEGOONN NIL)) VANAPH !2((OUTPUTF
(SSAY VANAPHDN))))
(M0540 NIL (ITALK) :(INPUTF (INTYPE ELLIP)) VELLIP !2((OUTPUTF ($SAY VELLIPDN)))
```

[GOON after LULL]

```
(M0570 NIL (ITALK) :(LULL (LEAD NIL) (INTRO NIL)) GOONN !2((OUTPUTF (SSAY GOONNDN))))
```

[Question inputs]

```
(M0600 NIL (ITALK) : (QUES) ANSWER)
(M0620 NIL (ITALK) : (QUES DISTRESSCH) PROVE ANSWER)
```

[LEADIN INTROTOP]

```
(M0650 NIL (ITALK) :(INPUTF (INTYPE TELLAB) (LEAD NIL)) !1((INTRO (GETINTRO #INPUTA))) INTROTOP)
(P0670 NIL (ITALK (TOPIC HOSPITAL)) (PINTERACTDN) NIL)
(P0671 NIL (ITALK PINTERACT (PINTERACTDN NIL) (DHOSTILE NIL)) ((WANTTOPIC @HOSPITAL) PINTERACTDN)
NIL)
(P0672 NIL (ITALK PHELP (PHELPCN NIL)) ((WANTTOPIC PHELP) PHELPCN) NIL)
(P0673 NIL ((OUTPUTF #1750)) (PHELPDN (PHELP NIL) (JOY 10)) NIL)
(P0674 NIL (ITALK PTELL (PTELLCN NIL)) ((WANTTOPIC PTELL) PTELLCN) NIL)
(P0675 NIL ((OUTPUTF #2110)) (PTELLDN (PHELP NIL) (JOY 10)) NIL)
(M0676 NIL (ITALK) :(LULL WANTTOPIC) !1((LEAD (FINDLEAD WANTTOPIC!)) LEADIN !2((WANTTOPIC (ANDV
(NOT TOPDONE) WANTTOPIC) )) )
(M0696 NIL (ITALK) :(PCONF LULL) VINTENT !1((SF FEELER)) SPECFN !2(PCONFDN (PCONF NIL)) )
```

[OPINION]

```
(M0730 NIL (ITALK) : (=INPUTA (INTYPE OPINION)) !1((TIACT @OPINION)) FINDOP .2((OUTPUTF (SSAY
FINDOPDN)) (TACT @OPINION) ))
```

[Clarifications]

```
(M0750 NIL (INCONV) :(QUES) (OUTPUTF) :(CLARQUES) (INPUTF) :(CLAR) !1((INPUTF SAVEINP)) ANSWER)
(P0760 NIL (QUES DEC) (CLARQUES) (VCLARIFQ) ((SAVEINP INPUTF)) )
(M0770 NIL :(OUTPUTF (OUTTYPE QUES!) (INPUTF) :(QUES) !1((SHAME 10)) )
```

[Finding and saying answers]

```
(M0810 ANSWER NIL :(INPUTF (LEADININP NIL)) !1(FFINDANS(FINDANS INPUTF)) FINDANS SAYANS
!2((OUTPUTF SAYANSDN) (TACT @A) ))
(P0830 NIL (FFINDANS (SHAME NIL)) ((TIACT @Q)) NIL)
```

[Finding and saying rejoinders]

```
(M0840 REJOIN NIL :(INPUTF) PROVE :((TALK I)) REJ SAYANS !2((OUTPUTF SAYANSDN)))
(M0860 REJ2 NIL :((TALK I)) REJ !2((OUTPUTF (SSAY REJDN))))
```

[Greetings]

```
(P0900 NIL (GREETINGS (=INPUTA HELLO)) NIL (PHELLO))
(P0902 NIL ((GREETINGS NIL) (=INPUTA HELLO)) !(SF @WEIRD)) NIL)
(P0903 NIL (GREETINGS (=INPUTA (HOW ARE YOU))) NIL (FINE))
(P0904 NIL (GREETINGS (=INPUTA =8150) (INTYPE OPINION)) NIL (FINE))
(P0906 NIL ((=INPUTA BYE)) ((B DQUIT)) (ENDCONV))
```

[General input questions]

```
(P0920 NIL (INPUTF (INTYPE YESNO)) (QUES) NIL (YESNO))
(P0921 NIL (INPUTF (INTYPE WHQUES)) (QUES) NIL (WHQUES))
(P0922 NIL (INPUTF (INTYPE ASSESS)) (QUES) NIL (ASSESS))
(P0923 NIL (INPUTF (INTYPE TELL)) (QUES) NIL (TELL))
(P0924 NIL (INPUTF (INTYPE WHICH)) (QUES) NIL (WHICH))
```

[Saying answers]

```
(M0850 SAYANS NIL VSAYANS !2 ((OUTPUTF SAYANSDN) (OUTTYPE @STMT)) )
```

[Types of rejoinders]

```
(M0950 REJ NIL :(AGREE (TRUTH TRUE)) GOONN)
(M0970 REJ NIL :(AGREE (TRUTH FALSE)) !1((ANGER 30) (B =DHONEST) (SF @ALOOF2)) SPECFN)
(M0990 REJ NIL :(AGREE (TRUTH NIL)) RESP)
(M1010 REJ NIL :(NOTAGREE NEGAFF) !1((SHAME 30) (SF @BELIEVEREPLIES)) SPECFN)
(M1030 REJ NIL :(NOTAGREE ORIG (NEGAFF NIL)) ORIGQ)
(M1050 REJ NIL :(NOTAGREE (ORIG NIL) (NEGAFF NIL)) ELABB)
(M1070 REJ NIL :((AGREE NIL) (NOTAGREE NIL) (TRUTH TRUE)) !1(VERIFY) GOONN)
(M1090 REJ NIL :((AGREE NIL) (NOTAGREE NIL) (TRUTH FALSE) (NEGAFF NIL)) !1(VERIFY) ELABB)
(M1110 REJ NIL :((AGREE NIL) (NOTAGREE NIL) (TRUTH FALSE) NEGAFF) !1((SHAME 30) VERIFY) RESP)
(M1130 REJ NIL :((AGREE NIL) (NOTAGREE NIL) (TRUTH NIL)) RESP)
```

[LEADIN and INTROTOP]

```
(M1170 INTROTOP NIL : (INTRO) VINTROTOP !2((NEWTOPIC INTRO) (OUTPUTF (SSAY VINTROTOPDN))) )
```

```
(M1191 LEADIN NIL :(LEAD) VLEADIN !2((OUTTOPIC LEAD) (DUM (ANAPHLEAD LEAD)) (OUTPUTF (SETQ
TOPDONE (SSAY VLEADINDN)) ) )
```

[Anaphs]

```
(P1220 NIL (GOONN) ((STORY (GETSTORY))) NIL)
(M1221 GOONN NIL :(STORY) VGOONN)
(P1241 NIL (GOON WANTTOPIC (STORY NIL) (LEAD NIL)) (LEADININP) NIL)
(M1242 GOONN NIL :((INTYPE ANAPH) (WANTTOPIC NIL) (STORY NIL)) !1(PUNT) )
(M1262 GOONN NIL :(YINPUTF (STORY NIL)) RESP)
(P1283 NIL (ELABB) ((ELABTN (GETANAPH @ELAB))) NIL)
(M1290 ELABB NIL :(ELABTN) VELABB)
```

[Get normal response to the (statement) input]

```
(P1321 NIL (RESP) ((RESP2 (RESPCHECK *INPUTF) )) NIL)
(M1322 RESP NIL :(RESP2) VRESP)
```

[Conditions for agreement]

```
(P1360 NIL ((*INPUTA (LIKE INT I))) (AGREE ILIKE) NIL)
(P1361 NIL ((*INPUTA (NOTLIKE INT I))) (NOTAGREE ILIKE) NIL)
(P1362 NIL ((*INPUTA YES)) (AGREE) NIL)
(P1363 NIL ((*INPUTA NO)) (NOTAGREE) NIL)
```

[Prove]

```
(M1370 PROVE NIL PROVES EXPLAIN !2 ((INPUTF (IMATCH XINPUTF)) (*INPUTA *INPUTF) PROVEDN REJ2 ))
(P1390 NIL (PROVES *INPUTA) ((OPPOS (OPPOS *INPUTA T))) NIL)
(M1391 PROVES NIL :(*INPUTA (TRUTH TRUE)) !1((ACTIV *INPUTA)) THINKS)
(M1411 PROVES NIL :(*INPUTA (INTYPE YESNO) (TRUTH FALSE)) !1((ACTIV *INPUTA) (ACTIV (OPPOS
*INPUTA T))) THINKS)
(M1431 PROVES NIL :(*INPUTA (INTYPE YESNO) OPPOS (TRUTH NIL)) !1((ACTIV *INPUTA)) THINKS
!1((SAVAFF AFF) (ACTIV (OPPOS *INPUTA T))) CHOOSEAFF !2((ACTIV (ADDEV (DBEL (OPPOS *INPUTF
OPBEL) )) )) )
(M1451 PROVES NIL :(*INPUTA (INTYPE YESNO) (OPPOS NIL) (TRUTH NIL)) !1((ACTIV *INPUTA)) THINKS)
(M1471 PROVES NIL :(INPUTF (INTYPE STMT) (TRUTH FALSE))
!1((ACTIV (ADDEV *INPUTA))) THINKS !2((SAVEAFF AFF))
!1((ACTIV (OPPOS *INPUTA T))) CHOOSEAFF
!2((ACTIV (ADDEV (DBEL (OPPOS *INPUTA OPBEL) )) )) )
(M1491 PROVES NIL :(*INPUTA (INTYPE STMT) (TRUTH NIL) OPPOS)
!1((SAVEAFF NIL) (ACTIV (ADDEV *INPUTA))) CHOOSEAFF !2((PROVESON (NOT OPBEL)) (SAVEAFF AFF))
:(OPBEL) !1((ACTIV (OPPOS *INPUTA T))) CHOOSEAFF2
!2((ACTIV (ADDEV (DBEL (OPPOS *INPUTA OPBEL) )) )) )
(M1511 PROVES NIL :(*INPUTA (INTYPE STMT) (TRUTH NIL) (OPPOS NIL)) !1((ACTIV (ADDEV *INPUTA)))
THINKS)
```

[Explain]

```
(M1540 EXPLAIN NIL :((INTYPE STMT) (TRUTH TRUE) POSAFF POSAFFIN) !1((SF @OPINION)) SPECFN)
(M1560 EXPLAIN NIL :((TRUTH TRUE) NEGAFF NEGAFFIN) !1((SF @ALOOF2)) SPECFN)
(M1580 EXPLAIN NIL :((TRUTH FALSE) (NEGAFF NIL) (NEGA  IN NIL)) !1((SF @FALSE)) SPECFN)
(M1600 EXPLAIN NIL :((TRUTH FALSE) NEGAFF (NEGAFFIN NIL)) !1((SF @ALOOF)) SPECFN,
(M1620 EXPLAIN NIL :((INTYPE STMT) (TRUTH FALSE) (NEGAFFIN NIL) NEGAFF) !1((SF @*DHONEST)) SPECFN)
(M1640 EXPLAIN NIL :(QUES (TRUTH FALSE) (NEGAFFIN NIL) NEGAFF) !1((SF @ALOOF)) SPECFN)
(M1660 EXPLAIN NIL :((TRUTH FALSE) NEGAFFIN POSAFF) !1((SF @ORIGO)) SPECFN)
(M1680 EXPLAIN NIL :((TRUTH FALSE) NEGAFFIN (POSAFF NIL)) !1((ANGER 30)) )
(M1700 EXPLAIN NIL :((INTYPE STMT) (TRUTH NIL) POSAFFIN NEGAFF) !1((SF @CAUTION)) SPECFN)
```

(M1720 EXPLAIN NIL VEXPLAIN)

[Setting intentions - done while waiting for next input]

(P3030 NIL (SAID (GDEC NIL)) (GOALS) (GDEC))

[Specific intentions]

(P3040 NIL (GOALS DDHELP (PHELPDN NIL)) ((PHELP (CHOOSELEAD))) NIL)
(P3041 NIL (GOALS PHELPDN DDKNOW DCHELP (WEAK NIL) (PTELLDN NIL)) ((PTELL @MAFIA)) NIL)
(P3042 NIL (GOALS PTELLDN DDKNOW (PCONFDN NIL)) (PCONF) NIL)
(P3043 NIL (GOALS (PHELPDN NIL) DDHELP PTELL) (PHELP (INTRO @MAFIA)) NIL)
(P3044 NIL (GOALS (ANGER HIGH) (PHARMDN NIL)) (PHARM) NIL)

Appendix C. Actions

A list of names and descriptions of the major actions in the system.

ENDCONV - halt the program

VEXIT - rate the interviewer and halt

VINSULT, VPANIC, VHOSTILE, VTHREATQ, VAFRAID, VDISTANCE, VSWEAR, VWKINSULT, QUESATTACK, ACCEPTAPOLOGY, VACCUSE, VALOOF, VSARCASM - reply with a special response

VEXHAUST - responds with a special set of exhaust replies

PUNTANS - reply with no knowledge of the input

VINTENT - set intentions

VSPECFN - special functions for remembering the doctor's name, getting the time of day

PHELLO - say hello

FINDANS - find the answer to a question

SAYCOM - output a comment

SAYQUES - output a question

SAYCMND - output an imperative

VINTROTOP - locate and introduce a new topic

VLEADIN - find the lead sentence for a story

FINDOP - form an opinion about an object or class

VGOONN - get the next line in a story

VANAPH - locate the referent for an anephoric input

VELLIP - locate the referent for an elliptical input

VELABB - elaborate on a topic

ORIGQ - ask why the interviewer mentioned a topic

VRESP - give a standard rejoinder to a statement

THINKS - test the truth value of an input

CHOOSEAFF, CHOOSEAFF2 - test the emotional reaction to an input

VEXPLAIN - select a number of responses based on input truth and emotional reaction

Appendix D. Beliefs

A list of the beliefs in the model. Beliefs have the following form:

(<belief name> <initial truth value>)

Annotations are in square brackets.

[Beliefs about Parry which lead to the paranoid mode if they become true]
(LYING 0) [Parry is dishonest]
(LOSER 0) [Parry is worthless]
(CRAZY 0) [Parry is crazy]
(DUMB 0) [Parry is dumb]

(CHEATB 0) [Parry cheated the bookie]
(NOTANS 0) [Parry isn't telling the whole truth]
(OBNOXIOUS 0) [Parry drives people away]
(NOFRIENDS 0) [Parry has no friends]
(NOCLASS 0) [Parry has no class, is a jerk]
(NOMONEY 0) [Parry has no money]
(BADJOB 0) [Parry has a bad job and can't get a better one]
(LOWSTATUS 0) [Parry comes from a family of low status]
(PARANOID 0) [Parry is paranoid]
(NEEDTREATMENT 0) [Parry needs special treatment]
(NEEDHOSP 0) [Parry needs to be in the hospital]
(STUPID 0) [Parry is stupid]
(BADSCHOOL 0) [Parry couldn't make it in school]
(NOTUNDERSTAND 0) [Parry doesn't understand the questions]

[Beliefs about the doctor's actions in conducting the interview]
(DCHELP 4) [doctor has the ability to help Parry]
(*DCHELP 2) [doctor does not have the ability to help Parry]
(DDHARM 2) [doctor wants to harm Parry]
(DDHARM 2) [doctor does not want to harm Parry]
(DDHELP 2) [doctor wants to help Parry]
(*DDHELP 2) [doctor does not want to help Parry]
(DDKNOW 2) [doctor wants to know more about Parry]
(*DDKNOW 2) [doctor does not want to know more about Parry]
(DDINTERACT 2) [doctor wants to interact with Parry]
(DQUIT 2) [doctor wants to stop the interview]
(DBEXCITED 2) [doctor believes Parry is excited]
(DBHELP 2) [doctor believes Parry wants help]
(DFACTS 2) [doctor asks factual questions]

[Belief about the doctor's traits]
(DSOCIABLE 4) [doctor is friendly]
(*DSOCIABLE 2) [doctor is not friendly]
(DTRUSTWORTHY 4) [doctor is trustworthy]

(*DTRUSTWORTHY 2) [doctor is not trustworthy]
(DRATIONAL 2) [doctor is rational]
(*DRATIONAL 2) [doctor is not rational]
(DHONEST 2) [doctor is honest]
(*DHONEST 2) [doctor is not honest]
(DHOSTILE 2) [doctor is hostile to Parry]
(DHELPFUL 2) [doctor is being helpful to Parry]
(*DHELPFUL 2) [doctor is not being helpful to Parry]
(DSIMILAR 2) [doctor has views similar to Parry's]
(*DSIMILAR 2) [doctor does not have views similar to Parry's]
(DCONFIDENT 2) [doctor is self-confident]
(*DCONFIDENT 2) [doctor is not self-confident]
(DDOMINATING 2) [doctor dominates the conversation]
(*DDOMINATING 2) [doctor does not dominate the conversation]
(DINITIATING 2) [doctor initiates subject areas and conversation paths]
(*DINITIATING 2) [doctor does not initiate subject areas and conversation paths]
(DGOOD 2) [doctor is competent]
(DMAFIA 2) [doctor has Mafia connections]
(DGANGSTER 2) [doctor is a gangster]
(DABNORMAL 2) [doctor is crazy]
(DEXCITED 2) [doctor is excited (angry, afraid, uptight)]

[Beliefs about the doctor's attitudes]
(DBNHONEST 0) [doctor believes Parry is dishonest]
(DBLOSER 0) [doctor believes Parry is worthless]
(DBABNORMAL 2) [doctor believes Parry is crazy]
(DBDUMB 0) [doctor believes Parry is dumb]
(DBELIEVE 2) [doctor believes Parry]
(*DBELIEVE 2) [doctor doesn't believe Parry]
(DDOCTOR 5) [interviewer is a doctor]
(*DDOCTOR 2) [interviewer is not a doctor]
(DGAMES 2) [doctor plays games]
(DBORED 2) [doctor is bored]
(DINSULTS 2) [doctor insults Parry]
(DSELFTALK 0) [doctor talks mostly about himself]
(DSELFFEEL 0) [doctor talks mostly about his own feelings]

[Beliefs about the interview]
(INTBAD 2) [the interview has been very bad so far]
(DBAD 2) [doctors in general are useless]

Appendix E. Rules of Inference

A list of the major rules of inference in the model. Annotations are in square brackets. The general form is:

(Pnnnn NIL PAST PRESENT NIL)

[Inferences about the interview]

(P3090 NIL ((B DDHARMX(DHELPFUL NIL)) ((B (INTBAD 3))) NIL)
(P3091 NIL ((B DQUITX(DHELPFUL NIL)) ((B (INTBAD 2))) NIL)
(P3092 NIL ((B DGAMESX(DHELPFUL NIL)) ((B (INTBAD 2))) NIL)
(P3093 NIL ((B DINSULTSX(DHELPFUL NIL)) ((B (INTBAD 2))) NIL)
(P3094 NIL ((B DHOSTILEX(DHELPFUL NIL)) ((B (INTBAD 2))) NIL)
(P3095 NIL ((B DBABNORMALX(DHELPFUL NIL)) ((B (INTBAD 4))) NIL)

[Inferences about paranoid beliefs]

(P3100 NIL ((B NEEDHOSP)) ((B CRAZY)) NIL)
(P3101 NIL ((B NEEDTREATMENT)) ((B CRAZY)) NIL)
(P3102 NIL ((B BADSCHOOL)) ((B DUMB)) NIL)
(P3103 NIL ((B NOTUNDERSTAND)) ((B DUMB)) NIL)
(P3104 NIL ((B BADJOB)) ((B LOSER)) NIL)
(P3105 NIL ((B LOWSTATUS)) ((B LOSER)) NIL)
(P3106 NIL ((B NOCLASS)) ((B LOSER)) NIL)
(P3107 NIL ((B NOFRIENDS)) ((B LOSER)) NIL)
(P3108 NIL ((B NOMONEY)) ((B LOSER)) NIL)
(P3109 NIL ((B OBNOXIOUS)) ((B LOSER)) NIL)
(P3110 NIL ((B CHEATB)) ((B LYING)) NIL)
(P3111 NIL ((B DNBELIEVE)) ((B LYING)) NIL)
(P3112 NIL ((B NOTANS)) ((B DUMBX(B DBNHONEST)) NIL)
(P3113 NIL ((*INPUTA (WISE REMARKS))) ((ANGER 10)) NIL)
(P3114 NIL ((B DBORED)) ((B *DDHELPX(B (LOSER 2))) NIL)

[Inferences about the interviewer's actions]

(P3116 NIL ((*INPUTA DDKNOW) SHAME) (PINSULT) NIL)
(P3117 NIL ((*INPUTA DDKNOW) JOY) (PCOMPLIMENT) NIL)
(P3118 NIL (INPUTF (INTOPIC MENTAL) DBABNORMAL) ((ANGER 40)) NIL)
(P3119 NIL ((B DNBELIEVE) (SHAME HIGH)) ((B DBABNORMAL)) NIL)
(P3120 NIL ((SHAME HIGHX(B SICK)) ((B CRAZY)) NIL)
(P3121 NIL ((*INPUTA CRAZY)) ((B DABNORMAL)) NIL)
(P3122 NIL (INPUTF (INTOPIC MENTALX(SHAME HIGH)) ((B DABNORMALX(FEAR 40))) NIL)
(P3123 NIL ((B DGANGSTER)) ((B DMAFIA)) NIL)
(P3124 NIL ((FEAR HIGH) (B DDHARMX(B DDOMINATING2)) ((B DGANGSTER)) NIL)
(P3125 NIL ((B DBNHONEST)) ((B DNBELIEVE)) NIL)
(P3126 NIL ((B (NOTAGREEWITH INT I))) ((B (DNBELIEVE 4))) NIL)
(P3127 NIL (INPUTF (INTOPIC GAMES)) ((B (DGAMES 3))) NIL)
(P3128 NIL (INPUTF (INTOPIC FACTS)) ((B (DFACTS 3))) NIL)
(P3129 NIL ((*INPUTA INSULT)) ((B DINSULTSX(ANGER 80)) NIL)
(P3130 NIL ((*INPUTA ATTACK)) ((B DINSULTSX(ANGER 80)) NIL)
(P3131 NIL ((*INPUTA *COMPLIMENT)) ((B DINSULTSX(ANGER 70)) NIL)
(P3132 NIL ((*INPUTA WINSULT)) ((B (DINSULTS 4)X(ANGER 30)) NIL)
(P3133 NIL ((*INPUTA CRAZY)) ((B DBABNORMALX(ANGER 40)X(FEAR 40)) NIL)
(P3134 NIL (INPUTF (INTOPIC YOU)) ((B (DSELFTALK 3))) NIL)
(P3135 NIL (INPUTF (SUBJ INT)) ((B (DSELFFEEL 3))) NIL)

(P3136 NIL ((*INPUTA (BE INT PRESIDENT))) ((B (DGAMES 5))) NIL)
(P3137 NIL ((*INPUTA (BE INT GOD))) ((B (DGAMES 5))) NIL)
(P3138 NIL ((*INPUTA (BE INT FAMILY))) ((B (DGAMES 5))) NIL)
(P3139 NIL ((*INPUTA $REPLIES)) ((B (DQUIT 2))) NIL)
(P3140 NIL ((*INPUTA $REPLIES)) ((B (DGAMES 3))) NIL)
(P3141 NIL ((*INPUTA (SWEAR INT))) ((B (DEXCITED 4))) NIL)
(P3142 NIL ((B (ANGRYAT INT I))) ((B DEXCITED)) NIL)
(P3143 NIL ((B (FEAR INT I))) ((B DEXCITED)) NIL)
(P3144 NIL (INPUTF (INTYPE IMPER)) ((B (DEXCITED 2))) NIL)
(P3145 NIL ((*INPUTA STHREAT)) ((B DDHARM)(FEAR 60)) NIL)
(P3146 NIL ((*INPUTA THREAT)) ((B DDHARM)(FEAR 50)) NIL)
(P3147 NIL ((*INPUTA (THREAT WEAK))) ((B (DDHARM 5)(FEAR 30)) NIL)
(P3148 NIL ((*INPUTA (HARM INT I))) ((B (DDHARM 5)(FEAR 30)) NIL)
(P3149 NIL ((B (KILL INT PEOPLE))) ((B (DDHARM 4))) NIL)
(P3150 NIL ((FEAR HIGH)(B DDOMINATING)) ((B (DDHARM 3))) NIL)
(P3151 NIL ((FEAR HIGH)(B DHOSTILE)(B DABNORMAL)) ((B DDHARM)) NIL)
(P3152 NIL ((*INPUTA (CHANGE I SUBJECT))) ((B (*DDKNOW 4))) NIL)
(P3153 NIL (INPUTF (INTOPIC IYCUME) (SUBJ I)) ((B DBEXCITED)) NIL)
(P3154 NIL ((B DDKNOW)) ((B DDINTERACT)) NIL)
(P3155 NIL ((B DSOCIABLE)) ((B DDINTERACT)) NIL)
(P3156 NIL ((*INPUTA (APPROVE INT I))) (GOODD) NIL)
(P3157 NIL ((*INPUTA (NOTAPPROVE INT I))) (BADD) NIL)
(P3158 NIL ((*INPUTA (AGREEWITH INT I))) (AGREE (B (DSIMILAR 4))) NIL)
(P3159 NIL ((*INPUTA (UNDERSTAND INT I))) (AGREE (B (DSIMILAR 4))) NIL)
(P3160 NIL ((*INPUTA (SYMPATHY))) (BADD) NIL)
(P3161 NIL ((*INPUTA (LIKE INT THAT))) (GOODD) NIL)
(P3162 NIL ((*INPUTA (NOTAGREEWITH INT I))) (NOTAGREE (B (*DSIMILAR 4))) NIL)
(P3163 NIL (POSAFF PLYING) (NOTAGREE (B (*DSIMILAR 4))) NIL)
(P3164 NIL ((POSAFF NIL) PLYING) ((ANGER 20) NOTAGREE (B (*DSIMILAR 4))) NIL)
(P3165 NIL ((*INPUTA COMPLIMENT)) ((B (*DHONEST 2)(B (DSOCIABLE 4)) PCOMPLIMENT) NIL)
(P3166 NIL ((*INPUTA (INSULT NOT))) ((B (*DHONEST 4)(B (DSOCIABLE 2))) NIL)
(P3167 NIL ((B DSIMILAR)) ((B DSOCIABLE)) NIL)
(P3168 NIL ((B *DSIMILAR)) ((B *DSOCIABLE)) NIL)
(P3169 NIL ((*INPUTA LYING)) (PLYING) NIL)
(P3170 NIL ((*INPUTA DNBELIEVE)) (PLYING) NIL)
(P3171 NIL ((FEAR HIGH)(DHELPFUL NIL)(B DINITIATING)) ((B (DHOSTILE 3))) NIL)
(P3172 NIL (INPUTF (ANGER MED) ANGERCH) ((B (DHOSTILE 4))) NIL)
(P3173 NIL (INPUTF (FEAR MED) FEARCH) ((B (DHOSTILE 5))) NIL)
(P3174 NIL (INPUTF JOY JOYCH) ((B (DHELPFUL 1))) NIL)
(P3175 NIL (INPUTF DDINTERACT (DISTRESS LOW)) ((JOY 30)) NIL)
(P3176 NIL ((B INTBAD)(B DGAMES)(B *DDHELP)) ((B *DDOCTOR)) NIL)

[Inferences about the interviewer's attitudes]

(P3177 NIL ((B *DRATIONAL)(B DHOSTILE)) ((B DABNORMAL)) NIL)
(P3178 NIL ((B DDHELP)(INTBAD NIL)(DABNORMAL NIL)(B DDOCTOR)) ((B DCHELP)) NIL)
(P3179 NIL ((B CRAZY)(B DMAFIA)) ((B DDHARM)) NIL)
(P3180 NIL ((B DDKNOW)(B DHELPFUL)(DDHARM NIL)(DBABNORMAL NIL)(DGAMES NIL)(DINSULTS NIL)) ((B DDHELP)) NIL)
(P3181 NIL (*INPUTA DDKNOW (DDHARM NIL)(DBABNORMAL NIL)(DGAMES NIL)(DINSULTS NIL)(ANGERCH NIL)(FEARCH NIL)) ((JOY 10)) NIL)
(P3182 NIL ((B DINITIATING)(DSELFTALK NIL)) ((B DDKNOW)) NIL)
(P3183 NIL ((B DDOMINATING)(DHOSTILE NIL)(DGAMES NIL)(DINSULTS NIL)) ((B DDKNOW)) NIL)
(P3184 NIL ((B *DINITIATING)(DINSULTS NIL)(DHOSTILE NIL)) ((B DDKNOW)) NIL)
(P3185 NIL (INPUTF NEWTOPIC) ((B (DDINTERACT 1))) NIL)
(P3186 NIL ((B DGAMES)) ((B (*DDKNOW 4))) NIL)
(P3187 NIL ((B DDHARM)) ((B *DDHELP)) NIL)
(P3188 NIL ((B DINSULTS)) ((B (*DDHELP 10))) NIL)
(P3189 NIL ((B DHOSTILE)) ((B (*DDHELP 7))) NIL)

(P3190 NIL ((B DBABNORMAL)) ((B (*DDHELP 5))) NIL)
(P3191 NIL ((ANGER HIGHXB DBABNORMAL)) ((B *DDHELP)) NIL)

[Inferences about the interviewer's traits]

(P3192 NIL ((B DEXCITED)) ((B (DDOMINATING 3))) NIL)
(P3193 NIL (INPUTF NEWTOPIC) ((B (DINITIATING 3))) NIL)
(P3194 NIL (INPUTF) ((B (DINITIATING 3))) NIL)
(P3195 NIL (INPUTF (INTYPE ANAPH)) ((B (*DINITIATING 5))) NIL)
(P3196 NIL (PUNT) ((B (*DINITIATING 3))) NIL)
(P3197 NIL (INPUTF NEWTOPIC) ((B (DDOMINATING 2))) NIL)
(P3198 NIL (INPUTF (INTYPE QUES)) ((B (DDOMINATING 2))) NIL)
(P3199 NIL ((B DSELFTALKXB DINITIATING)) ((B DDOMINATING)) NIL)
(P3200 NIL ((B *DINITIATINGXDHOSTILE NILXDEXCITED NILXDGAMES NIL)) ((B *DDOMINATING)) NIL)
(P3201 NIL ((B DHOSTILE)) ((B (*DHELPFUL 6))) NIL)
(P3202 NIL ((B DDHARM)) ((B *DHELPFUL)) NIL)
(P3203 NIL ((B DGAMES)) ((B (*DHELPFUL 5))) NIL)
(P3204 NIL ((B DDKNOWXDHOSTILE NIL)) ((B (DHELPFUL 3))) NIL)
(P3205 NIL ((B DINSULTS)) ((B (*DHELPFUL 4))) NIL)
(P3206 NIL (INPUTF (PHELP NIL) (INTOPIC MAFIA)) ((B (DMAFIA 5))) NIL)
(P3207 NIL ((B DSIMILARXB DDINTERACT)) ((B DHONEST)) NIL)
(P3208 NIL ((ANGER HIGH) (B DHOSTILE)(B DDKNOW)) ((B *DHONEST)) NIL)
(P3209 NIL ((B DDINTERACTXB DSELFTALK)) ((B (DSOCIABLE 4))) NIL)
(P3210 NIL ((B DHOSTILE)) ((B *DSOCIABLE)) NIL)
(P3211 NIL ((B DDHELP)) ((B (DTRUSTWORTHY 4))) NIL)
(P3212 NIL ((B DHONEST)) ((B (DTRUSTWORTHY 4))) NIL)
(P3213 NIL ((B DHOSTILE)) ((B *DTRUSTWORTHY)) NIL)
(P3214 NIL ((B DDHARM)) ((B *DTRUSTWORTHY)) NIL)
(P3215 NIL ((B *DHELPFUL)) ((B *DTRUSTWORTHY)) NIL)
(P3216 NIL (PUNT) ((B (*DRATIONAL 2))) NIL)
(P3217 NIL ((B DGAMESXB DNBELIEVE)) ((B *DRATIONAL)) NIL)
(P3218 NIL ((B DSIMILARXB DHONESTXDHOSTILE NIL)) ((B DRATIONAL)) NIL)
(P3219 NIL ((B DDOMINATINGXB DSELFTALK)) ((B DCONFIDENT)) NIL)
(P3220 NIL ((B *DINITIATINGXB DQUIT)) ((B *DCONFIDENT)) NIL)
(P3221 NIL ((B *DINITIATINGXB DGAMES)) ((B *DCONFIDENT)) NIL)
(P3222 NIL ((B DHOSTILEXDINITIATING NIL)) ((B *DCONFIDENT)) NIL)

Bibliography

TINLAP is *Theoretical Issues in Natural Language Processing*. R. C. Schank and B. L. Nash-Webber (Eds.), Cambridge, Mass., 1975.

IJCAI-3 is *Proceedings of the Third International Joint Conference on Artificial Intelligence*, Stanford, Cal., 1973.

IJCAI-4 is *Proceedings of the Fourth International Joint Conference on Artificial Intelligence*, Tbilisi, USSR, 1975.

IJCAI 5 is *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, Cambridge, Mass, 1977.

Abelson, R. P. Computer simulation of "hot" cognition. In S. S. Tomkins & S. Messick (Eds.), *Computer Simulation of Personality*. New York: Wiley, 1963.

Abelson, R. P. Concepts for representing mundane reality in plans. In Bobrow and Collins, 1975.

Apter, M. J. *Computer Simulation of Behaviour*. London: Hutchinson & Co., 1970.

Arnold, M. B. *Emotion and Personality*. New York: Columbia University Press, 1960.

Averill, J. R., Opton, E. M., Jr., & Lazarus, R. S. Cross-cultural studies of psychophysiological responses during stress and emotion. *International Journal of Psychology*. 4, 83-102, 1969.

Becker, J. D. A model for the encoding of experiential information. In Schank & Colby, 1973.

Bobrow, D. G. & Collins A. (eds.), *Representation and Understanding: Studies in Cognitive Science*. San Francisco: Academic Press, 1975.

Boden, M. A. Intentionality and physical systems. *Philosophy of Science*, 37, 200-214, 1970.

Boden, M. A. *Purposive Explanation in Psychology*. Cambridge: Harvard University Press, 1972.

Boden, M. A. The structure of intentions. *Journal for the Theory of Social Behavior*. 3, 1, 23-46, 1973.

Boden, M. A. Freudian mechanisms of defense: A programming perspective. In R. Wohlheim (Ed.), *Freud, A collection of critical essays*. Garden City, New York: Anchor Press, 1974.

Bruce, B. C. Belief systems and language understanding. BBN Report No. 2973, BBN, Inc., Boston, Mass, 1975a.

Bruce, B. C. Generation as a social action. In TINLAP, 1975b.

Bruce, B. C. Pragmatics in speech understanding. In IJCAI-4, 1975c.

Carnap, R. *Meaning and Necessity*. Chicago: University, 1947.

Charniak, E. Organization and inference in a frame-like system of common knowledge. In TINLAP, 1975.

Chisholm, R. M. *Perceiving: A Philosophical Study*. New York: Cornell U. Press, 1957.

Clippinger, J. H. A discourse speaking program as a preliminary theory of discourse behavior and a limited theory of psychoanalytic discourse. Dissertation, University of Pennsylvania, 1974.

Clippinger, J. H. Speaking with many tongues: Some problems in modeling speakers of actual discourse. In TINLAP, 1975.

Colby, K. M. Experimental treatment of neurotic computer programs. *Archives of General Psychiatry*, 10, 220-227, 1964.

Colby, K. M., Weber, S., & Hilf, F. D. Artificial paranoia. *Artificial Intelligence* 2, 1-25, 1971.

Colby, K. M. *Artificial Paranoia*. New York: Pergamon Press, 1975.

Colby, K. M. Clinical implications of a simulation model of paranoid processes. *Archives of General Psychiatry*, 1976.

Colby, K. M. An appraisal of four psychological theories of paranoid phenomena. *J. Abnormal Psych.* 86, 1, 54-59, 1977.

Colby, K. M., Hilf, F. D., Weber, S., and Kraemer, H. Turing-like indistinguishibility tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence*. 3, 199-221, 1972.

Colby, K. M., Parkison, R. C., and Faught, W. S. Pattern-matching rules for the recognition of natural language dialogue expressions. *Am. J. Computational Linguistics*. Microfiche 5, Sept. 1974.

Craik, K. J. W. *The nature of explanation*. Cambridge: Cambridge University Press, 1943.

Dennett, D. C. *Content and Consciousness*. New York: The Humanities Press, 1969.

Deutsch, B. G. The structure of task oriented dialogs. IEEE Symposium on Speech Recognition, 1974.

Doran, J. E. Experiments with a pleasure-seeking automaton. *Machine Intelligence 3*, pp. 119-35 (eds Dale, E. & Michie, D.). Edinburgh: Edinburgh University Press, 1968.

Doran, J. E. Planning and generalization in an automaton/environment system. *Machine Intelligence 4*, pp. 433-54 (eds Meltzer, P. & Michie, D.). Edinburgh: Edinburgh University Press. 1969.

Erdelyi, M. H. A new look at the new look: Perceptual defense and vigilance. *Psych. Review*. 81, 1, 1-25, 1974.

Faught, W. S. Affect as motivation for cognitive and conative processes. In IJCAI-4, 1975.

Faught, W. S., Colby, K. M., and Parkison, R. C. Inferences, affects, and intentions in a model of paranoia. *Cognitive Psychology*, 9, 153-187, 1977.

Fikes, R. E., Hart, P. E. & Nilsson, N. J. Learning and executing generalised robot plans. *Artificial Intelligence*, 3, 251-288, 1972.

Fodor, J. A. *The Language of Thought*. New York: Crowell, 1975.

Frege, G. Ueber Sinn und Bedeutung. *Zeitschrift fuer Philosophie und philosophische Kritik*. 100, 25-50, 1892. English translation (On sense and reference) in G. Frege, *Philosophical Writings*. (P. Geach and M. Black, eds.) Oxford: Blackwell, 1952.

Hayes-Roth, F. and Mostow, D. J. An automatically compilable recognition network for structured patterns. In IJCAI-4, 1975.

Heiser, J. F., Colby, K. M., Faught, W. S., and Parkison, R. C. Testing Turing's test: Can psychiatrists distinguish a computer simulation of paranoia from the real thing? UCLA Algorithmic Laboratory of Higher Mental Functions, Memo ALHMF-12, July, 1977.

Izard, C. E. *The Face of Emotion*. New York: Appleton Century Crofts, 1971.

Kilmer, W. L., McCulloch, W. S., & Blum, J. A model of the vertebrate central command system. *Int. J. man-mach. Studies*, 1, 279-309, 1969.

Kiss, G. R. Outlines of a computer model of motivation. In IJCAI-3, 1973.

Leeper, R. W. A motivational theory of emotion to replace "emotion as disorganized response." *Psychological Review*. 55, 5-21, 1948.

Loehlin, J. C. *Computer Models of Personality*. New York: Random House, 1968.

McDougall, W. *An Introduction to Social Psychology*. London: Methuen, 1908.

Miller, G. A., Galanter, E., & Pribram, K. H. *Plans and the Structure of Behavior*. New York: Holt, Rinehart, and Winston, 1960.

Minsky, M. L., A framework for representing knowledge. In Winston (ed.), *The Psychology of Computer Vision*. New York: McGraw-Hill, 1975.

Mischel, T. (Ed.) *Human Action*. New York: Academic Press, 1969.

Mischler, E. G. Studies in dialog and discourse. *J. Psycholing Res.* 4, 2, 99-121, 1975.

Moore, J. and Newell, A. How can MERLIN understand? In *Knowledge and Cognition*. (Ed. Gregg, L.), New York: Wiley & Sons, 1974.

Newell, A. & Simon, H. A. GPS: A program that simulates human thought. In E. A. Feigenbaum and J. Feldman (Eds.) *Computers and Thought*. New York: Mcgraw-Hill, 1963.

Newell, A., and Simon, H. *Human Problem Solving*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Norman, D. A. & Rumelhart, D. E. & the LNR Research Group. *Explorations in Cognition*. San Francisco: W. H. Freeman, 1975.

Parkison, R. C., Colby, K. M., and Faught, W. S. Conversational language comprehension using integrated pattern-matching and parsing. *Artificial Intelligence*, 1977.

Plutchik, R. *The Emotions: Facts, Theories, and a New Model*. New York: Random House, 1962.

Quine, W. V. O. *Word and Object*. Cambridge, Mass.: MIT Press, 1960.

Rieger, C. J. On organization of knowledge for problem solving and language comprehension. *Artificial Intelligence*, 2, 89-127, 1976.

Ryle, G. *The Concept of Mind*. London: Hutchinson, 1949.

Sacerdoti, E. D. The nonlinear nature of plans. In IJCAI-4, 1975.

Sacerdoti, E. D. A structure for plans and behavior. Dissertation, Stanford University, 1975.

Schachter, S. The interaction of cognitive and physiological determinants of emotional state. In *Anxiety and Behavior* (ed. C. D. Spielberger). New York: Academic Press, 1966.

Schachter, S. and Singer, J. Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, 69, 379-99, 1962.

Schank, R. C. & Abelson, R. P. Scripts, plans, and knowledge. In IJCAI-4, 1974.

Schank, R. C. & Colby, K. M. (Eds.) *Computer Models of Thought and Language*. San Francisco: W. H. Freeman, 1973.

Schatzman, M. Paranoia or persecution: The case of Schreber. *Family Process*, 10, 2, 177-212, 1971.

Schmidt, C. F. & D'Addamio, J. A model of the common-sense theory of intention and personal causation. In IJCAI-3, 1973.

Schmidt, C. F. & Sridharan, N. S. The plan recognition problem: A hypothesize and revise paradigm. In IJCAI-5, 1977.

Scragg, G. W. A structure for actions. Working paper no. 20, Castagnola, Switzerland: Institute for Semantic and Cognitive Studies, 1975.

Shaw, D. E., A strategy for making computers understand. *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, San Francisco, 1975.

Simon, H. A. Motivational and emotional controls of cognition. *Psychological Review*, 74, 29-39, 1967.

Simon, H. A. *The Sciences of the Artificial*. Cambridge, Mass.: MIT Press, 1969.

Solley, C. M. & Murphy, G. *Development of the Perceptual World*. New York: Basic Books, 1960.

Spiegel, L. A. Affects in relation to self and object. *The Psychoanalytic Study of the Child*. 21, 69-92, 1966.

Sridharan, N. S. & Schmidt, C. F. Knowledge-directed inference in Believer. In *Proc. Workshop on Pattern-Directed Inference*. May, 1977.

Suppes, P. and Warren, H. On the generation and classification of defence mechanisms. *Int. J. Psycho-Anal.*, 56, 405-414, 1975.

Swanson, D. W., Bohnert, P. J., & Smith, J. A. *The Paranoid*. New York: Little, Brown, & Co., 1970.

Tomkins, S. S. *Affect, Imagery, and Consciousness*. New York: Springer, 1962.

Waterman, D. A. Adaptive production systems. In IJCAI-4, 1975.

Winograd, T., Frames and the declarative-procedural controversy. In Bobrow and Collins, 1975.

Winograd, T. A framework for understanding discourse. In *Cognitive Processes in Comprehension*, (Carpenter and Just, eds.), Lawrence Erlbaum Associates, 1977.

Wright, G. H. von, *Explanation and Understanding*. Ithaca, N. Y.: Cornell U. Press, 1971.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER (14) STAN-CS-77-633, AIM-305	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle) (6) Motivation and Intensionality in a Computer Simulation Model,	5. TYPE OF REPORT & PERIOD COVERED (9) Doctoral Thesis,
6. PERFORMING ORG. REPORT NUMBER AIM-305	
7. AUTHOR(s) (10) William Simmons Faught	8. CONTRACT OR GRANT NUMBER(s) (15) MDA903-76-C-0206, PHS-NIMH-06645-13
9. PERFORMING ORGANIZATION NAME AND ADDRESS Artificial Intelligence Laboratory Stanford University Stanford, California 94305	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS LPR ARPA Order 2494
11. CONTROLLING OFFICE NAME AND ADDRESS Eugene Stubbs ARPA/PM 1400 Wilson Blvd., Arlington, VA 22209	12. REPORT DATE (11) Sep 77
	13. NUMBER OF PAGES 104 (12) 107p.
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Philip Surra, ONR Representative Durand Aeronautics Building, Room 165 Stanford University Stanford, California 94305	15. SECURITY CLASS. (of this report) 15
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Releasable without limitations on dissemination

DISTRIBUTION STATEMENT A
Approved for public release; Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This dissertation describes a computer simulation model of paranoia. The model mimics the behavior of a patient participating in a psychiatric interview by answering questions, introducing its own topics, and responding to negatively-valued (e.g., threatening or shame-producing) situations.

The focus of this work is on the motivation mechanisms required to instigate and direct the modelled behavior.

The major components of the model are:

(1) A production system (PS) formalism accounting for the instigation and guidance of behavior as a function of internal (affective) and external (real-world) environmental factors. Each rule in the PS is either an action pattern (AP) or an interpretation pattern(IP). Both may have either affect (emotion) conditions, external variables, or

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

12. Cont'd

outputs of other patterns as their initial conditions (left-hand sides). The PS activates all rules whose left-hand sides are true, selects the one with the highest affect, and performs the action specified by the right-hand side.

(2) A model of affects (emotions) as an anticipation mechanism based on a small number of basic pain-pleasure factors. Primary activation (raising an affects's strength) occurs when the particular condition for the affect is anticipated (e.g., anticipation of pain for the fear affect). Secondary activation occurs when an internal construct (AP, IP, belief) is used and its associated affect is processed.

(3) A formalism for intensional behavior (directed by internal models) requiring a dual representation of symbol and concept. An intensional object (belief) can be accessed either by sensing it in the environment (concept) or by its name (token). Similarly, an intensional action (intention) can be specified either by its conditions in the immediate environment (concept) or by its name (token).

Issues of intelligence, psychopathological modelling, and artificial intelligence programming are discussed. The paranoid phenomenon is found to be explainable as an extremely skewed use of normal processes. Applications of these constructs are found to be useful in AI programs dealing with error recovery, incompletely specified input data, and natural language specification of tasks to perform.