A MODEL OF THE COMMON-SENSE THEORY OF INTENTION AND PERSONAL CAUSATION*

Charles F. Schmidt
Psychology Department
Rutgers University
New Brunswick, New Jersey
and
John D'Addamio
Computer Science Department
Rutgers University
New Brunswick, New Jersey

Abstract

Certain general properties of man's ability to interpret the actions of other persons are discussed. Some distinguishing features of this common-sense theory include the nature of the modal operators of Can and Try, the asymmetry of implication, and the capacity to embed models within models.

The structure of a proposed model of this naive theory of personal causation is presented. This model arrives at a specific interpretation of another's actions by showing that these actions represent a possible path to a particular goal that is consistent with the axioms of the belief system's theory of human motivation and personality organization.

Descriptors

belief systems, intention, personal causation, problem solving

Introduction

The goal Of the research to be discussed here is to develop a model of how persons can arrive at an interpretation of the social actions of other persons. By interpretation is meant the layman's ability to determine the reasons that might have motivated another person's action. This question of how man determines the intentions of another person's actions is of central importance to the area of social psychology. One person's interpretation of the actions of another almost invariably affects the way in which the observer reacts to and evaluates the actor. This process of naive intentional analysis is also crucially involved in communication situations. The psychiatric interview and diplomatic exchanges are quasi-ritualized examples of communication situations where the importance of intentional interpretations is dramatically present.

Within the intersection of artificial intelligence and cognitive psychology some notable work has been begun on developing a model of how a listener might arrive at a conceptual representation of an utterance-'-" as well as work investigating the way in which incoming information is integrated by a human belief systemS,4,5,6, Some linguists have also begun to give increasing attention to the role that presupposition plays in human language and discourse?.9)12. However, very little attention has been given to the explication of the logic that underlies man's capacity to infer the intentional basis for another's action. It is hoped that the present work will begin to fill this gap in our approach to the investigation of the processes involved in human communication.

Working within the area of social psychology Heider[8] has provided an extraordinarily Tich analysis of social perception and a number of theorists have provided important extensions and reinterpretations of this analysis'' * However, the absence of an adequate formalism which is both rich and explicit enough to do justice to these theoretical schemas has hampered the development of this area. The purpose of the present paper is to demonstrate the usefulness of employing certain ideas drawn from work in artificial intelligence in developing a theory of social perception.

Before proceeding to the specific manner in which we have attempted to account for this type of social understanding, it will be useful to briefly state the major psychological properties that must he incorporated within a theory of social understanding.

The basic assumption of man's common-sense theory of human action is that man has choice. Without this assumption the notion of intention would be superfluous and the concept of personal causation would be no different than that of non-personal causation. However, personal causation is not a concept that can be defined independently of the common-sense concept of physical causation. The person is acting in a physical system and is therefore constrained by this system. In order to explain the why of choice, this naive theory must be capable of determining when choices are available, which particular choices are possible, and what outcomes follow as necessary or possible consequents of a particular action that could be chosen. Thus there is a relation between these two concepts of causality and the answer to the question of why a person chose to act in a particular way involves showing how the particular act or act sequence chosen might or did lead to the achievement of the actor's goal.

This interrelation between the physical and the personal causative systems is reflected in Heider's suggestion that personal causation involves two sets of conditions which he labels the Can and the Try conditions. Using these terms the basic postdictive axiom of personal causation may be written as

Cause Cp,w) ---> CanCp.w) A Try (p,uO

where w represents either an act, e.g. Give (p_T,p,,,x) or an outcome condition, e.g. Get (p.,x). This axiom states that if the naive observer believes that a person, p, intentionally caused some act or outcome, then the observer will infer that the person was able to do the act or cause the outcome and was motivated to do the act or cause the outcome. The specific reason that motivated the actor is simply a particular proof of Try *(p,w)* that the observer's belief system can derive from the axioms, theorems, and data that are currently present in the observer's belief system. These Try axioms represent the assumptions of the observer's common sense psychological theory.

In addition to the capacity to postdictively interpret the actions of others, we are also at times

capable of determining when an actor's plan has gone awry, of recognizing that a sequence of acts is only a part of a larger plan that is yet to be completed, and also of making judgements concerning the set of alternative courses of action that were available to an actor. The existence of these abilities implies that man is capable of projecting possible action sequences. However, this projective capacity is only weakly predictive. To reflect this fact, the projective axiom of personal causation is written as

$$\text{Can } (p,\omega) \wedge \text{Try } (p,\omega)\ p \supset \text{Cause } (p,\omega)$$

where $p\supset$ is to be read as 'implies the possibility'. This asymmetry of postdictive and projective implication is one of the important distinguishing characteristics of the common sense theory of social action. Projective implication is assumed to be weak for two reasons. First, the axioms of the Try component include reference to the internal psychological states of the actor and the observer cannot directly observe these states. Secondly and most basically, the common sense theory of action does not assume that psychological states determine a person's choice of action - psychological states only set the occasion for action.

These axioms that comprise the Try component are thoroughly psychological and serve to specify the classes of particular goals that an actor can be expected to pursue at a particular moment. An example of a rule of this type would be the common-sense notion of hedonism, that is, the belief that persons initiate actions that they believe will result in outcomes beneficial to themselves. The Can component, however, includes the common-sense knowledge of the physical world and, in this sense, this aspect of the lay theory is conceptually similar to the concept of can as developed by McCarthy[13]. However, the Can component as utilized in the logic of personal causation also has psychological aspects connected with it.

One such aspect results from the fact that the psychological preconditions or presuppositions associated with beliefs concerning rules of action include preconditions on the actor as well as preconditions on the environment. For example, in order for a person to be able to fly a plane requires not only that a plane be available but also that the person have the necessary knowledge and ability to fly a plane.

Psychological aspects enter into this Can component in another very significant way. Often a particular person, A, cannot directly cause some particular outcome. However, either through the assistance of another or through the agency of another person, B, the actor, A, may be able to cause the desired outcome. Situations such as this are common in our social activities and it is this type of situation that points up an expecially interesting aspect of the common-sense theory of personal causation. Whenever one person must involve another person in his plan for achieving some goal the person must be able to insure that both Can and Try are true for that other person. Conversely, whenever a person acts to achieve his own goal that person must at times insure that Can or Try are false for a person who might be motivated to block his acts and thus thwart his plan. This aspect of the interpretation of other's actions greatly complicates the common-sense theory of personal causation. This aspect implies that the actor, A, must be capable of interpreting the intentions and possible intentions of other actors with whom he must deal. Therefore, in order to interpret this actor's actions the observer must be capable of interpreting the situation from A's point of view. Of course, A may utilize his ability to view his own contemplated action from his representation of B's point of view and use this information in dealing with B, and so on. The point is simply that our common-sense theory possesses the capacity to embed models within models and this fact coupled with the modalities introduced by the Can and Try operators yields a system of enormous subtlety and complexity. How deeply our common-sense theory typically embeds models within models and how broad a range of possible actions is typically explored is unclear. However, it is clear that these capacities exist and any model of this common-sense theory of personal causation must include these properties.

These considerations represent the basic structural assumptions that have guided the development of the specific model described below. The immediate goal of this work is to create a computer-implemented model which will accept descriptions of actions and provide as output the reasons why the persons acted in the manner described. Additionally, the model should be empirically credible -- that is, these model-generated reasons or explanations should be indistinguishable from explanations provided by human subjects who have been given the same descriptions to explain. Although the model described below has been partially implemented and is currently capable of postdictively explaining action sequences, the stage of systematic empirical test of this model has not yet been reached. The purpose of this report is to communicate the general outline of how such a model of this aspect of human behavior can be formalized. The intent is not to claim psychological reality for each assumption embodied within the model.

Proposed Model

Human action occurs within a context and our interpretation of other's actions is context dependent. Therefore, the appropriate unit of analysis is not the isolated act, but rather a sequence of actions occurring within a situational context. It is assumed that the observer must have the following type of information available in order to interpret the actions of another person. First, the observer must have partial knowledge of the properties that hold just prior to the initial action. This set is termed the initial situation (S_0). Secondly, the observer must have knowledge of the ordered sequence of actions (A_i) that have occurred. This represents the information that must be minimally available. In addition, the observer may also know the properties that hold in the situation at intermediate points (S_i) in the action sequence. Finally, if the observer is previously acquainted with the actors, then the observer's belief system may contain a set of existing beliefs (B) about the properties of the actors.

This collection of information represents the input to the model. This input may be symbolically represented as:

$$\text{Input} = [(B),\ S_0,\ A_1,\ (S_1),\ A_2(S_2),\ldots,A_i(S_i)\ldots]$$

where the parentheses indicate that the information is optionally present. This collection of information has been rather vacuously labelled the input in order to emphasize that it is not necessarily an episode. An episode is a psychologically structured entity. One of the tasks of the model is to discover one or more episodes that may be completely or partially present in the input.

Overview of the Model

It has been assumed that this ability to structure and interpret social actions implies that the observer is acting as a kind of extended problem solving system. A typical problem solving system is given a goal and then attempts to find a sequence of legal moves that transform the initial state into a state which satisfies the goal conditions. The observer of social actions assumes that the actors are engaged in problem solving activity — that they have goals -- but the observer does not typically know what specific goals are being pursued. Therefore, the observer must possess a set of rules, termed Try rules, which implicitly specify the set of psychologically permissable goals. This set of Try rules corresponds to the axioms of the Try component of the axiom of personal causation.

These Try rules are of two different types. The first type are termed the General Try rules. These rules represent the axioms of human motivation that the observer believes to be applicable to all persons. The hedonism rule mentioned previously is an example of a rule of this type. The second type of Try rules are termed Dispositional Try rules. In general, these latter rules are applied only if the General Try rules fail. These dispositional rules are the rules which are used to make inferences concerning the personality of the actors, and these rules are the source of the naive theory of personality. An example will help to clarify this distinction. Most persons that we know act honestly most of the time. Nonetheless, very few of us would characterize the majority of persons whom we know as possessing the trait 'honest'. Thus, acting honestly is not in and of itself a sufficient basis for inferring that a person is honest. However, if a person acts honestly in a situation where he can gain nothing from acting honestly and would gain much and lose nothing from acting dishonestly, then that person is honest. In terms of Try rules, the person broke a hedonism rule in acting honestly and it is this deviation from the General Try rules that creates the occasion for a Dispositional rule to apply.

The identification of a permissable goal in the situation is not sufficient. The goal identified must fit the data provided by the actions. That is, the observer must be able to establish that an action or action sequence represents a set of moves that are consistent with the attainment of the goal state specified by a particular Try rule. This implies that the naive theory must possess a set of rules of action. These action rules must specify not only the conditions that must hold in the situation in order for the action to be taken, but also the set of outcome conditions that may result if the action is taken. Using these rules, the observer's interpretive system is capable of determining whether the Can component of the axiom of personal causation is true. It is this aspect of the model that is analogous to a problem solving system.

By using the Can condition in conjunction with the Try rules a unique interpretation of a sequence of actions can often be obtained. However, at times several interpretations are possible and the situation is ambiguous. If the observer possesses some previously inferred beliefs about the actors, these beliefs are often utilized to help resolve the ambiguity. We have termed this set of rules the Consistency rules. The basic notion that these rules are meant to capture is that if the observer already has inferred that the actor possesses a certain disposition, then interpretations that are inconsistent with the existing dispositional beliefs are avoided. A mean, helpful person or an honest, dishonest person does not seem to be psychologically believable. The fact that this type of personality configuration is in some sense anomalous provides a basis for the argument that the naive theory must possess these consistency rules which function as a kind of meta-theory of personality organization. If these rules are violated we tend to look deeper into the situation in search of a more acceptable interpretation. Thus, this set of rules acts as a kind of monitor and filter of the output of the Can and Try rules.

Representation of Actions, Persons and Situations

Natural language is the major system of expression used to inform others of intentions, to discuss plans, and to tell stories. Gesture, facial expression, posture and intonation are also used to communicate intention and emotion, but the subtleties of these systems of expression are beyond the scope of our model. For these reasons, we have looked to natural language for an appropriate set of concepts and properties for representing actions, persons and situations. The underlying assumption that has been adopted is that lexical items describing properties of persons and verbs describing interpersonal actions are related in a systematic fashion and that much of the logic of personal causation is implicit in our language.

It is useful to think of the verbs of English as falling into two general classes. The first class specifies either the existence of a property of an entity or a relation between entities. Examples of verbs used to communicate the existence of properties of persons are: has, owns, knows, believes, is able to, wants, and needs. Verbs specifying the existence of relations between persons are exemplified by the verbs: married to, son of, friend of, likes, hates, and dominates. The second class of verbs are those which can be used to describe interpersonal actions. These include verbs usually employed to describe the exchange of physical objects (e.g. give, take, buy, sell, steal) as well as verbs used to describe the exchange of information (e.g. tell, ask, command, insult, beg, and threaten). Aside from these actions of exchange there are also verbs used to denote movement, ingestion, and so on. However, the exchange acts appear to be the most crucially involved in describing social actions. For this reason, it is this class of action verbs that has been given the most attention.

The verbs specifying properties and relations and those denoting action are related in an interesting way. Verbs of action generally presuppose the existence of certain properties in the situation. For example, give presupposes that the actor intends the action and that the actor possesses the object which is being given. Furthermore, each action verb carries with it some information concerning the consequences that necessarily or possibly result after the action has been taken. This observation supports the assumption that underlying each action verb is knowledge of an action schema which defines a partial function consisting of a set of antecedent properties that must hold immediately prior to the occurrence of the act and a set of consequent properties that may hold after the action has been taken. Stated more formally, an act schema is represented as:

Name Act $[E_1.E_2...)$ $\stackrel{*}{\leftarrow}$ (Try Flag) * {PC} / (0C)

where (0C) ■ {NOC} u {POC). The action schema consists

on the left side of the name of the act together with a set of dummy arguments ($E_1, E_2, ,.$) which restrict the class of entities that can be substituted as actor, recipient, object, and instrument of the act. The right side includes the Try Flag whose value is non-nil if the action implies intention on the part of the actor, the set {PC} which contains the list of preconditions on the act, and *the set {OC}* which designates the set of outcome conditions associated with the action. The slash is used to visually segregate the preconditions from the outcome conditions. The set of outcome conditions is partitioned into the set that necessarily occur if the action is performed {NOC}, and those that possibly happen, {POC}, if the action is performed. These pre- and post-conditions consist of properties and relations that may hold in the situation in which the action occurs. Thus, the function of the first class of verbs is to partially specify the state of the world and the latter class, the action verbs, serve as the operators that can transform the present situation into a future situation.

A particular person is represented as a list structure headed by the person's name. This list structure is differentiated into the general property classes listed in Figure 1. Beneath each class are examples of particular properties of that class. This structure is motivated by the rules of the belief system rather than by any a-priori psychological hypothesis. Except for the disposition and wants classes, each of these properties can enter as members of the set of pre- and post-conditions associated with an act schema. What happens to a person simply involves either getting or losing properties belonging to one of these seven classes. The knowledge and belief classes are used to differentiate what is cognitively available from what is believed.

In addition to the name of each particular property the belief system associates a value with each property. This value represents the system's belief concerning the person's evaluation of the property. Two sources of value are distinguished. The first is the property's intrinsic value and represents man and society's general assumptions about what is good, bad or indifferent. This value is static and relatively context free. The second source of value is dynamic and context sensitive and is termed pragmatic value. This is the value that a property takes on by virtue of its role in enabling or blocking a particular outcome. For example, the intrinsic Value of possessing a Club foot is probably negative, but if the existence of this property can block *a* young man from being drafted then that young man may, for this reason, value this property quite highly. In general, the pragmatic value depends on the value of the outcome that is enabled or blocked.

The want properties represent the belief system's list of goals that are thought to be currently relevant to the person. These goals are used primarily as a source of candidate goals when the system is attempting to draw projective implications.

The dispositional properties are also related to the operation of the Try component. A disposition is associated with a person if a particular Dispositional Try rule was used to explain some previous action of the person. A dispositional property such as helpful or honest is the memory tag for this previous inference about the person. Associated with a dispositional property is a set of coordinates giving the location of the disposition in an implicational space of disposition terms. These coordinates are

Person Name

(1) Biological States:
e.g. sick, hungry, healthy, etc.

(2) Emotional States:
e.g. happy, sad, angry, etc.

(3) Abilities:
e.g. able to perform surgery, able to play chess, etc.

(4) Possessions:
e.g. has $2000, has Ferrari, etc.

(5) Knowledges:
e.g. knows Mary has a headache, knows auto mechanic claims the crankshaft must be replaced, etc.

(6) Beliefs:
e.g. believes Mary has a headache, believes the crankshaft doesn't need replacing, etc.

(7) Interpersonal Relationships:
(a) Unit Relations
e.g. person married to Mary, person father of Tom, etc.

(b) Sentiment Relations
e.g. person loves Mary, person hates Sam, etc.

(c) Dominance Relations
e.g. person boss of Larry, person employee of Sam, etc.

(8) Dispositions
e.g. is helpful, is honest

(9) Wants
e.g. wants to be friend of Susie, etc.

Figure 1. Properties of Persons

used by the Consistency rules to maintain a permissable personality structure.

The situation or environment also consists of a list of properties. These properties are used in an ad hoc fashion and consist of the environmental conditions that must be known by the belief system in order to interpret a particular input. For example, in order for a medical doctor to operate on a patient the appropriate medical environment must be available. Rather than attempting to systematize all of the environmental information that might be needed this information is created when it is necessary.

Input Representation

The input of a situation and set of actions is represented as a list consisting of three major sublists corresponding to persons, environmental conditions, and actions. The list of persons is composed of the names of all persons involved in the story together with the list of properties that are believed to hold for each of these persons at the start of the story. The environmental conditions are likewise a list of all of the environmental properties that are believed to exist at the start of the story. This

information is either provided by the narrator or it is retrieved from a file where permanent properties are stored. These two sublists represent the initial situation, S_0.

Using the set of act schemas a fully instantiated and ordered tree of the actions and their pre- and post-conditions is created. This action tree together with S_0 represent all of the information provided as input. The action tree is then searched for action dependencies. This is done by starting with the last action, act , and searching to determine whether any of the preconditions for act were created by some previous action, act , . If so, then a link is created connecting act , and act . If no prior enabling act is found then'it is assumed that the precondition existed at the start of the story and a link is created pointing back to the initial situation S_0. This search for enabling connections is carried out over all previous acts for all preconditions on each act.

Two kinds of inferences are made as a result of this process. The first involves previously unknown and unspecified properties of persons or the environment. If, for example, one act was that 'John drove a car to the airport', then the belief system could infer from this that John possessed the ability to drive a car. This type of inference, while rather mundane, is the source of our ability to fill in a large amount of the information that is implicit in a story. A second and more interesting type of inference involves the distinction between the types of outcomes that are associated with actions, namely, these that necessarily follow and those that possibly follow from the action. If an outcome that is represented as a possible outcome of some act. is found to be a precondition for a subsequent action, act., and this condition is not known to have held previous to act., then the system infers that this possible outcome actually happened and changes its representation accordingly. This type of inference is particularly important for inferring a person's beliefs. For example, if a medical doctor tells John that his wife must have an operation we know that John believes that the doctor believes that the operation is necessary. However, John may or may not believe in the necessity of the operation. However, if John hires the doctor to operate on his wife then it is possible to infer that in fact John also believes that the operation is necessary and the doctor's act of telling was actually an act of convincing.

In this way the action dependency function fills out the story and also narrows down the potential set of outcomes. By locating action dependencies that might exist between the actions, this function provides a partially structured representation that is interpreted by the Try rules.

Try Rules

Four types of General Try rules and the Dispositional Try rules have been developed. The General Try rules include hedonism, extended hedonism, reciprocity and normative rules. Hedonism is the common-sense notion that actions may be taken because outcomes beneficial to the actor are expected to result. Extended hedonism generalizes this notion to include the possibility that the actor expects the outcomes to benefit some person connected to the actor by a positive unit or sentiment relation, for example a son or a friend. The core idea of reciprocity is that persons are expected to respond in kind to the actions of others. Normative rules represent beliefs about cultural or legal norms that apply to a particular situation and usually include a belief about an action that some 'norm enforcer' will try to take if the norm is broken. A situationally instantiated local normative rule can also be established if one person makes a statement such as "If you don't hand over your money, then I will kill you". The Dispositional rules generally apply only if certain General Try rules are broken. This corresponds to the intuitive notion that a person's actions that deviate from the general rules of action provide the most information about that person.

These are all relatively mundane notions. However, the interesting point is that coupled with the act rules and action dependencies these rules can be generalized in very powerful ways. These Try rules are implemented as a kind of grammar which parses the action structure. The terminal strings of these rewrite rules are actually functions defined on the action data base rather than elements and their order is irrelevant. A Try rule is said to explain a set of actions if each of these functions returns the value of True. A greatly simplified fragment of the hedonism component of this grammar can be developed to exemplify this approach.

A and B will denote particular persons and X a particular object that has been instantiated in the actions of the story. Then the top-level rule may be written as

$$\text{Try}(A,act_i) \rightarrow \text{Hedonism}_i \vee \text{Ext. Hedonism}_i \vee \text{Reciprocity}_i \vee \text{Normative}_i \vee \text{Disposition}_i$$

where i is the ordinal position of the particular act in question and j is greater than i. The Hedonism fragment is expanded as

$$\text{Hedonism}_i \rightarrow \text{Immediate}_i \vee \text{Possible}_i \vee \text{Self-Enabling}_i \vee \text{Other-Enabling}_i$$

$$\text{Immediate}_i \rightarrow [\text{Get}(A,X) \in \{HOC_i\} \wedge V(X_A) = \langle + \rangle] \vee [\text{Lose}(A,X) \in \{HOC_i\} \wedge V(X_A) = \langle - \rangle]$$

$$\text{Possible}_i \rightarrow [\text{Get}(A,X) \in \{POC_i\} \wedge V(X_A) = \langle + \rangle] \vee [\text{Lose}(A,X) \in \{POC_i\} \wedge V(X_A) = \langle - \rangle]$$

$$\text{Self-Enabling}_i \rightarrow [\exists y \in \{OC_i\} \mid (\sim \text{Has}(S_{i-1},y) \wedge \text{fact}_j \mid (y \in \{PC_j\} \wedge \text{Cause}(A,act_j) \supset (\text{Immediate}_j \vee \text{Possible}_j \vee \text{Self-Enabling}_j \vee \text{Other-Enabling}_j)))]$$

$$\text{Other-Enabling}_i \rightarrow [\exists y \in \{OC_i\} \mid (\sim \text{Has}(S_{i-1},y) \wedge \text{fact}_j \mid (y \in \{PC_j\} \wedge \text{Cause}(B,act_j) \supset (\text{Immediate}_j \vee \text{Possible}_j \vee \text{Self-Enabling}_j \vee \text{Other-Enabling}_j)))]$$

where $\{HOC_i\}$ represents the set of outcomes of act_i that are known to have happened and $V(X_A)$ refers to the value of X with respect to person A. The sequence of acts created by self-enabling or other-enabling is only topologically ordered.

This example shows that the simple notion of hedonism has been extended in two ways. First, the rule can be satisfied in possible worlds as well as in the world that actually resulted. Secondly, some of the rules are defined recursively and therefore an entire act sequence may be explained by a single application of the hedonism rules. In this way a person can be located as a rather remote causal source of a sequence of actions.

Extended hedonism is developed in the same fashion as hedonism except that the outcome is associated with someone with whom the actor is related by a positive sentiment or unit relation. For example, to help a friend would satisfy extended hedonism.

The principle of reciprocity can also be extended in a very powerful way. However, a very tedious development of notation is needed to express the ideas formally. Therefore, a brief and informal sketch will *suffice.* The reciprocity condition is that if some previous action of B, act., has some ye{OC.} that could or did affect A and[1]A believes that B intended the act, then if A does some act. such that some ze{0C.} can or does affect B in a[j] similar way, then A's action is explained by reciprocity. Again, this notion can be extended over possible worlds and can also be extended recursively in a fashion analogous to the hedonism rules. Therefore, reciprocity can be used to explain even very indirect connections such as preemptive strike where A hits B because A believes that B's act is part of an enabling sequence that will allow B to hit A. Reciprocity is also extended in a way analogous to extended hedonism. That is, if B hurts A's friend, then A may hurt B or a friend of B and so on.

The disposition rules also possess some interesting properties. The English language is filled with dispositional *concepts,* for example, helpful, exploitative, etc. The interesting property of these terms is that many of the terms are derived rather directly from verbs or actions. This suggests that some of the action rules may be used as the basis for dispositional Try rules. However, these Try rules can apply only when certain of the general Try rules are broken. To demonstrate this approach, the disposition rule for 'helpful' will be developed. The act rule for 'help' is defined as

$$\text{Help}(A,B,\text{act}_j(A,\ldots)) \rightarrow \text{Try}(A,\text{Help}) \wedge \text{Know}(A,\text{Want}(B,X)) \wedge [\exists \text{act}_j \mid \text{Can}(A,\text{act}_j) \wedge [\text{Cause}\{A,\text{act}_j\} \supset [\text{Get}(B,X) \in \{OC_j\} \vee SH_j]]] \text{ where } SH_j = [\exists y \in \{OC_j\} \mid \text{Can}(B,\text{act}_k) \wedge [\text{Get}(B,X) \in \{OC_k\} \vee SH_k]] \;/\; \text{Get}(B,X)$$

The preconditions of the help rule provide the basis for the Try rule, Helpful. A Helpful Try rule is obtained by adding the conditions that Hedonism, Extended Hedonism, Reciprocity, and Normative rules are broken.

Consistency Rules

A dispositional interpretation is checked by the Consistency rules. This check involves computing the implicational distance between the existing dispositional properties and the disposition that is the basis for the particular Dispositional rule that has been found true. If this distance exceeds a specified value, this interpretation is rejected and the Try rules must search deeper for an alternative interpretation.

Models within Models and Projective Implication

The system is capable of interpreting the actions of others from the point-of-view of a particular actor by restricting its beliefs to the beliefs that are represented for that particular actor. The reciprocity example above included the condition that A believe that B intended act.. This condition involves a recursive call on thi Try rules where Try of B is computed from A's point-of-view.

This capacity to determine interpretations of actions from a particular actor's point-of-view is also involved whenever projective implications are made. There are two classes of conditions that commonly involve projection. The first occurs whenever alternative choices of action are considered. For example, in an episode where John has stolon some money which enables his wife to have a needed operation, an extended hedonism rule is satisfied but a normative rule is broken. The resolution of this ambiguity may depend upon whether or *not* from John's point-of-view it can be shown that the Can of alternative courses of action, such as attempting to borrow the money, are true. The second case involving projective implication occurs when the system attempts to complete a partial plan. For example, if all that is known is that John's wife is in need of an operation and that John has stolen some money, then the system should be capable of determining whether or not John's action may actually be an enabling action in a larger causal sequence.

Attempts to project the implications of an action sequence often involve a search over a very large set of possible goals and possible paths. Therefore, this capacity must be used by man only under a narrow range of conditions and the system must possess a powerful set of heuristics to guide and limit this search. The nature of these heuristics is *still very much* in question. Abelson[1] in his work on the hierarchical structure of belief systems has suggested that plans may be organized into themes and these themes themselves may be elements of a larger structure termed scripts. This is a very intriguing suggestion and perhaps the incorporation of more complex structures of this type will provide the means for efficiently projecting plans. At present, our model is clearly deficient in this regard.

References

1. Abelson, R. P. The structure of belief systems. In K. Colby & R. Schank (Eds.), Computer simulation of thought and language. Freeman, in press, 1973.

2. Abelson, R. P. Psychological implication. In R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, & P. H. Tannenbaum (Eds.), Theories of cognitive consistency: A sourcebook. Chicago: Rand McNally, 1968.

3. Abelson, R. P., & Carroll, J. D. Computer simulation of individual belief systems. American Behavioral Scientist, 1965, 8, 24-30.

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

5. Colby, K. M. Computer simulation of change in personal belief systems. Behavioral Science, 1967, 12, 248-253.

6. Colby, K. M., Tesler, L. & Enea, H. Experiments with a search algorithm on the data base of a human belief structure. Stanford Artificial Intelligence Project, Memo No. AI-94, 1969.

7. Gordon, D., & Lakoff, G. Conversational postulates. In Papers from the seventh regional meeting of the Chicago Linguistic Society. Chicago Linguistic Society, 1971.

8. Heider, F. The psychology of interpersonal relations. New York: Wiley, 1958.

9. Hutchinson, L. G. Presupposition and belief inferences. In Papers from the seventh regional meeting of the Chicago Linguistic Society, '' ~~ Chicago Linguistic Society, 1971.

10. Jones, E. E., § Davis, K. E. From acts to dispositions. The attribution process in person perception. In L. Berkowitz (Ed.), Advances in experimental social psychology. Vol. II. New' York: Academic Press, 1965.

11. Kelley, H, H. Attribution theory in social psychology. In D. Levine (Ed.), Nebraska symposium on motivation. Lincoln: University of Nebraska Press, 1967.

12. Lakoff, G. Presupposition and relative well-formedness. D. D. Steinberg & L. A. Jakobovits (Eds.), Semantics. Cambridge: Cambridge University Press, 1971.

13. McCarthy, J. Programs with common sense. In M. Minsky (Ed.), Semantic information processing. Cambridge, Mass.: MIT Press, 1968.

14. Schank, R. C. Conceptual dependency: A theory of natural language understanding. Cognitive Psychology, 1972, 3, 552-631.

This work was supported by NIH grant RR-643.