MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

Working Paper 269 April 1985

A Proposal for Research With the Goal of Formulating A Computational Theory of Rational Action

by John Batali

Abstract:
A theory of rational action can be used to determine the right action to perform in a situation. I will develop a theory of rational action in which an agent has access to an explicit theory of rationality. The agent makes use of this theory when it chooses its actions, including the actions involved in determining how to apply the theory. The intentional states of the agent are realized in states and processes of its physical body. The body of the agent is a computational entity whose operations are under the control of a program. The agent has full access to that program and controls its actions by manipulating that program. I will illustrate the theory by implementing a system which simulates the actions a rational agent takes in various situations.

This document has been submitted to the department of Electrical Engineering and Computer Science at MIT as a "Proposal for Thesis Research in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy." I thank the following people for helpful comments and suggestions: Phil Agre, Mike Brady, David Chapman, Gary Drescher, Margaret Fleck, David Kirsh, Ron Rivest, Gerald Jay Sussman, Susan Wiegand.

CONTENTS

1. Rationality

An agent is situated in the world. The world is composed of physical objects and changes through time. Events occur and are related to one another. The agent is a being of the world -- it is a physical object and hence can participate in the world. The agent can affect the world according to its abilities, and can be affected by the world. The agent's relationship with the world is *Intentional*: some things *matter* to the agent, some things in the world *mean* things to the agent, the agent has goals and desires and intentions and beliefs, all of which are *about* things or states of affairs in the world.

The agent can perform *actions*. As a physical object in the world, changes in the agent can cause changes in the world. As an Intentional being, the agent can perform actions in the world in order to realize certain effects. Since actions are supposed to have certain results, we can speak of success or failure of actions. We can speak of the agent taking an action to prepare for other actions, or to achieve a certain goal or to satisfy a desire. Actions, in general, depend on the situation both in order to enable their being taken, and for the effects they will have.

Given an agent with certain abilities and certain goals, intentions, or desires, in a specific situation, we can judge if a certain action is reasonable or sensible, or that one action would be better or worse than another. An agent in the course of acting considers its situation and its abilities, chooses an action to take, and then attempts it. This "practical reasoning" also involves a judgment as to the best or right action for an agent (in this case the one making the judgment) in a situation. I suggest that the term *rational* is applied to actions which are judged to be correct, right, or good ones to perform in a situation. Judgments about the rationality of an action are *defeasible* in that more information about the situation, agent, or proposed action, can change judgments about the rationality of the action.

Rationality is the notion of the "right action" to do in a situation. A theory of rationality would be a theory which could be applied to a situation and an agent in order to ascertain the quality of certain potential actions in that situation. As we are rational agents, the theory would be *prescriptive* in that it would indicate the right thing for any rational agent to do in that situation: the results of the theory take the form "a rational agent (and therefore you) *ought* to do this action in this situation." But the theory will also involve a *description* of the workings of a rational agent. In particular, the rational agent is one which performs (or tends to perform) the rational act in a situation *because* it is the

rational act according to the prescriptive theory. A theory of rationality must explain how the rational agent is bound by, and obeys, the prescriptions of the theory.

Rationality is not perfection. While there may be a better thing to do in some situation, there may not be a *better thing for the agent to do,* given the abilities of the agent and particulars about its situation. Rationality involves dealing with the *finiteness* of the agent, with its faults, inadequacies and limitations. An important way in which agents are finite is in their ability to judge the rationality of actions. Humans are rational agents to the extent that they attempt to "do their best" in their day to day activities. But I will not develop any notion of "perfect" rationality to which humans are mere approximations. The notion of rationality only makes sense for finite, limited agents.

Dealing as it does with questions of "ought" and "should" and "right", a theory of rationality has a close relation with ethics. Indeed most ethical theories have presupposed a rational agent who attempts to be (or just is) also a moral agent. I will be examining the question of the rightness of actions in terms of the actions themselves, not with respect to any conception of morality. I will be dealing with a *solitary* agent -- one which is alone in a world with no other rational agents -- so the ethical questions of relating to other rational agents will not be addressed.

2. Motivation for this Research

Artificial intelligence needs a theory of rational action. To the extent that artificial intelligence is about *intelligence*, the field really is the study of rationality. But very little research in AI broaches the subject directly. Much of AI research has concentrated on limited domains, or on specific abilities. In the first case, the expertise of a program in a particular domain need bear no relationship to the question of rationality in general. In the second case, the importance of a specific ability (*eg:* finding analogies), or a choice of details in a theory thereof, is very difficult to judge in the absence of a more general framework into which the ability is ultimately supposed to fit.

Not only must AI devote attention to general theories of rationality, but it must also shed the dualistic tradition which holds that the seat of rationality is a disembodied mind. The attraction of this view (apart from its having been the more-or-less standard view of Anglo-American philosophy) is that it suggests a natural role for computation. In modern Cognitive Science, mental states and processes are specified in computational terms. These states and processes are then assumed to be implemented in the brain, which is viewed as a kind of computer. But it is only the abstract processes and states which constitute the "mental" aspects of the theories, interaction with the world is limited to formally specifiable input and output modules. This makes the most fundamental aspect of the "mental states" -- their being *about* things "outside" the mind -- the source of the most unwieldy problems such theories must ultimately resolve.

Instead I suggest that we theorize about embodied *agents* situated in the world. There is only the world, and the agent is a part of it. Instead of "mental" entities, we should speak of the characteristically "Intentional" aspects of those physical entities which are agents. The Intentional states of the agent (for example, its having a certain desire) are realized in physical states of the agent's body. The difference between actions that affect the world outside the agent's body and those that affect states of the agent is merely a difference in the location and specifics of physical effects, not a difference between "mental" and physical actions. Thinking is a kind of doing.

There is a natural place for computation in theories of this kind also. Computers are physical objects in the world. A computer program is a representation of the behavior of a physical object. A token of a program, placed in a specific relation to a computer can cause the very behavior it represents. Computation is therefore a link between the Intentional world of representations and the physical world of states and events. A computational theory of rationality would involve an account of the

relationship between representations of actions and their being taken. And the "realized in" relationship between the Intentional states of a rational agent and the physical states of its body could be the reasonably well understood "implemented in" relationship between a computational system and its hardware and software components.

Artificial intelligence work has always presupposed some idea of rationality. The use of a heuristic in a program involves the feeling that the heuristic action is reasonable or justified or sensible in a particular situation. A specific representational system is presented with certain assumptions about the sort of useful inference actions which must be made with it. And all AI researchers have had to make decisions about control. This information, most often fixed by the programmer, embodies a more or less explicit commitment that some sequence of steps is "better" or "more reasonable" than its alternatives.

AI systems often solve problems involving complex decisions by means of a search through a formalized space. In most interesting domains, searches are difficult and expensive, and often are not even guaranteed to produce the desired result. So the best thing to do in such domains is not to search at all. Instead it would be superior to think about what is being done in a way that makes the search unnecessary, or to reformalize the search space to make the problem easier. This means that programs must have much greater access to their control structures, both to allow the programmer to specify more clever control strategies, and to allow a program to manipulate its control structure to best advantage. Programs must become programmers; their domains of expertise must include themselves. Programs must be made to understand the implicit theory of "best action" -- of *rationality* -- which programmers use when designing them. This means that a theory of rationality must be made explicit and the explicit theory must be made accessible to programs.

3. Outline of the Theory

The theory of rational action which I am developing is based on viewing thinking as a kind of action and thus itself subject to the prescriptions of the theory. The theory is *intellectualist* in that it depicts rationality as a matter of explicit reasoning. A rational agent is viewed as being *introspective* in that all of its internal states (for example its intentions and desires) are subject to inspection and manipulation by the agent. It is by this manipulation that the agent controls its actions. The theory is *computational* in that the rational agent is viewed as physically embodied in an object which may be controlled by a representation of the desired behavior: a program.

The theory has two main parts: the "structural theory" which describes what an agent *can* do, and the "deontic theory" which describes what an agent *should* do. The theory is *recursive*, in that the generation of a rational act requires the agent to perform other rational actions as precursors (for example: deliberating about what to do). In the following paragraphs I will present a terse summary of the theory as it stands. The full development and completion of the theory remains a goal of this research.

3.1 The Structural Theory

The structural theory is the part of the theory of rationality which deals with the states and events and processes involved in the generation of rational acts. In the structural theory we describe the agent as being in a situation and as performing the rational act for that situation. One of the things an agent can do in any situation is to choose to deliberate about the situation and apply the theory of rationality in order to determine what to do. So the structural theory must include these deliberation actions as a part of the structure of situations and actions surrounding an action.

To express and represent the structural theory, we must first have a model of how events occur in the world. A "model of occurrences" is an account of how events occur and are related. The model involves "intervals" which are partial descriptions of the world. An interval is "satisfied" if the world is as described for a certain period of time. Intervals are related in time, and events in intervals can causally affect events in other intervals. The world contains physical objects and the behavior of those objects can be described by describing the possible sequences of intervals satisfied by the object over time.

The world also contains the rational agent. An ability involves a "possible event" which the agent may cause. The agent must be able to interpret physical objects as representing other objects or states of affairs. The agent may have *intentions* which are states that represent and cause actions of the agent. And, as I have emphasized, some of the actions of the agent are "internal" in that they affect only the state of the agent; for example, the action of coming to have a specific intention.

Within the model of occurrences we will be able to express the notion of a plan or an action having a goal or purpose: roughly speaking, the goal of an action is the occurrence (or class of occurrences) which would count as the action's "succeeding." We can characterize the notion of planning as a process of actions (representable as occurrences involving the agent) which attempts to discover a representation of a process which will satisfy a certain set of goals and which is within the abilities of the agent in some possible situation. The execution of a plan will involve steps taken to test the world to make sure that the plan is succeeding.

I will represent the structural theory of rationality within this model of the world and the agent. As the agent is presumed to deal with an explicit theory of rationality, the structural theory must include descriptions of the theory itself, and the agent's use of it in determining what to do. And since the agent must work with these descriptions, there must be an explicit theory of *representation* which the agent can use to understand what it is doing as it manipulates these descriptions.

3.2 The Deontic Theory

The deontic theory of rationality is used to decide what the agent ought to do. It is the aspect of the theory which actually tells the agent what to do, or how to figure out what to do. These conclusions, called "rational imperatives", are expressed in the structural theory, as are the situation and the potential actions of the agent. A rational imperative specifies a rational action for a specific situation. Some rational imperatives are general policies explicitly formulated and represented, perhaps while the agent is idly thinking. Other imperatives are specific to the situation at hand and are expressed in the action of the agent.

There are very few situations for which we can specify the correct action for all agents, for all time. And judgments of rationality are such that in many cases, new information about, or even a new description of a situation may change the adjudged rationality of an action in that situation. Taking this into account, what is the deontic theory to prescribe?

As one answer consider the following line of reasoning, which is expressible in the structural theory: If the agent has a representation of its past actions and the situations in which they were performed, and the agent confronts a new situation which is like a previous situation in some way, then it might be rational to perform an action like the action previously taken. Why is this the rational act to take? Because the agent represents itself as a rational agent, and thus as performing the rational act in a situation. Since it remembers that it performed a specific act in a situation past, it can therefore conclude that that action was the rational action to take in that situation. And so it should take a similar action now. "Precedential Reasoning" is the term I give to this notion of deciding what to do on the basis of similarities with past situations. Most of the rational imperatives in the theory I am proposing will be based on precedential reasoning.

Precedential reasoning involves finding analogies between the current situation and previous ones. This will involve finding mappings in the structural theory between the situations. From *any* action the agent has done, it may conclude the rational imperative to do *that* action in *that* situation. It may then generalize that imperative by mapping the specific structure of the action to less specific structures, applicable to more situations. The agent gains expertise in a domain by solving particular problems in the domain. Since the agent is also reasoning about its reasoning actions, it will gain expertise in the domain of reasoning about actions. In some cases it might be able to generalize across domains as, for example, when a similar control decision occurs in different domains.

Another class of rational imperatives determined by the deontic theory are consequences of the finiteness of the agent. In general this class is negative, in the sense that it allows the agent to conclude that certain actions are *not* rational. From "ought implies can", the agent can conclude that "can not implies ought not". Some such limitations involve physical abilities of the agent -- an agent which cannot weld should not make plans which require it to do so. An important limitation is that of time: when decisions are time-critical the agent must often commit to an action which more lengthy deliberation would reject. Other limitations involve "internal" abilities or knowledge of the agent. For example if an agent is aware of only one method to solve a certain class of problems, then that method is the one it ought to use, even if other methods exist.

While the limitations of the agent provide it with some answers to deontic problems, they also provide the agent with more problems it must deal with. For example, as the agent is not perfect, it will make mistakes and must determine what went wrong and attempt to correct them. This fact undermines the justification for precedential reasoning -- the previous action might not have been rational. Still the

previous action is probably the best suggestion for what to do now, especially if it seemed to work the previous time. But the agent must be aware of the fact that this conclusion might be mistaken, and in general, must always be ready to deal with failures and unexpected results.

3.3 The Body is a Computer

The striking aspect of practical reasoning is that it is linked to action. When an agent concludes that an action is rational, it takes that action (at least it often does). I call this intimate connection between the judgment and the action *elan* -- the property of a rational agent to actually do what it determines to be rational. To Aristotle, elan was the most significant feature of practical reasoning, which he referred to as reasoning "whose conclusion is an action". Indeed in the present theory, practical reasoning is just like "theoretical" reasoning about the right action except for elan: The methods of reasoning are the same and the same conclusions are justified in the same way but in practical reasoning the agent does something when the conclusion is reached.

To understand elan, we recall that the body of the agent is a physical object in the world. And the Intentional states of the agent are realized in physical states and processes of the body. Deliberation, like other actions of the agent, involves physical processes. So the connection between deliberations and action is just physical causation. But more must be said: the action which is taken is taken *because* it is rational. The physical events occurring in the body of the agent respect the theory of rationality. They do so because the agent, as a rational agent, *makes* them do so. The rational agent controls its body by setting it up so that its states and processes will realize the appropriate Intentional states for rational action.

This leads to a view of the body of the agent as a *computational* entity. A computer is a physical object whose behavior is under the causal control of a representation of that behavior. One controls a computer by manipulating its program. So the rational agent controls its actions by manipulating its program. The agent controls the physical states and processes in which its Intentionality is realized by *implementing* those states and processes computationally. Recall that the theory of rationality specifies what the agent should do in situations. As the agent is rational, this is what it *does* do, and it does so by consulting an explicit theory of rational action. To consult the theory, the agent must have a representation of the theory available. This representation of the theory of rationality stands in the same relation to the body of the agent as a program does to a computer -- the theory describes what

the agent will do, and also causally controls what the agent will do. So the theory of rationality is part of the program for the rational agent.

The main way that the agent controls its actions by manipulating its program is to set up its body so that it will perform the rational action in specific situations. To do so, it must first obtain a rational imperative from the deontic theory. It will then examine its program and determine the state the program will enter when the situation described by the imperative occurs. It will then modify the program so that this state is followed by the attempt to perform the action specified by the imperative. Later, when the agent's body enters this state, the agent need do no deliberation -- it will just perform the action. I call this kind of manipulation of the agent's program by the agent the "wiring-in" of an action. Wired-in actions are rational -- they are performed because they are the rational thing to do. But the determination of their rationality for the situation is performed before the situation occurs. Most wiring-in causes a *class* of actions to be taken because the situation in which the action is wired-in may be specified to any degree of generality. By a single action of wiring-in, the agent can set itself up to perform many rational acts "automatically" in the future.

Most of the actions a rational agent takes are wired-in. This is possible because most situations are enough like previous situations that precedential reasoning can be used to determine the appropriate actions. So although the agent is fully introspective in that it *can* access any aspect of its program, it rarely does so. It only needs to explicitly deal with a situation if it has no wired-in response to that situation, or if its wired-in response fails. Since many actions of the agent are wired-in there may be cases where a number of actions should and can be performed simultaneously. This may be the case especially for internal actions, for example, considering a large number of possible actions in parallel with simple wired-in criteria, and selecting a small number of promising actions for more careful non-wired-in deliberation.

Agents may very well make mistakes and wire-in actions which cause behavior which, in specific instances, could be judged to not be rational. However this does not contradict the rationality of the agent. The action of wiring-in *was* rational, given what the agent knew then, or given how much time the agent had to worry about how perfect to make the wiring-in. In some sense, by wiring-in actions, the agent has limited itself to perform the wired-in action in the future. When the agent later enters the relevant situation, it performs the "irrational" action because it *must* -- because it is wired-in. So the action isn't really irrational in the situation where it was performed because the agent had no choice but to do it. Of course, the performance of irrational-seeming actions by the agent is a hint to

it that it has made mistakes in its wiring-in and ought to "unwire" some actions or otherwise fix its program.

3.4 Rationality by Recursion

The two kinds of solutions to problems from the deontic theory require that the agent view itself as, respectively, a rational agent and a finite agent. Both solutions involve deliberation by the agent about itself and its situation. Deliberation is a kind of action, so deliberation is also something to which the theory of rationality is applicable. There is a "right thing" to do when deliberating. Deliberative actions must be representable in the structural theory, and solutions to deliberative "what should I do?" problems must be available from the deontic theory.

The theory is developed by elucidating these problems in the structural theory. The rational agent must solve these subproblems in the generation of a rational act. Some such subproblems the agent must solve include: finding a mapping between a previous situation and the current one; determining that one of its limitations applies to the present situation; estimating how much time is available for deliberation; discovering that some previous action was a mistake and how to fix it. Each of these problems may be just as hard as any of the others, including the "main" problem which is being considered. Yet each must be solved.

For there to be solutions to these problems the agent must have access to enough previous situations so that a precedential case will be found to apply and/or it must have enough information about its limitations (and enough limitations) to make decisions about what it must do in specific situations. I call this the "deontic grounding" which is required for the theory of rationality to work -- there must be enough problems solved so that other problems *can* be solved.

Furthermore, enough of these situations must have wired-in actions so that the agent will be able to perform some non wired-in action. If *no* actions were wired-in at all, the agent could never act at all because it would need to somehow decide what to do, but this decision is itself an act which would have to be chosen and so on. (This regress, by the way, is one of the major objections to the intellectualist approach to rationality. See, for example, [Dreyfus] and [Ryle].) There must be some wired-in actions for enough classes of situations so that this regress is terminated. There must, that is, be some actions which *just happen* so that the agent need not deliberate about them. I call this the "elan grounding" for the theory.

So the theory of rationality is recursive in that it is defined as involving actions which themselves require other rational actions to take place. The "body" of the recursion is represented in the structural theory which describes the various actions and how they are related. The base cases of the recursion come from a large set of particular situations: the "deontic grounding" is a set of specific precedents and limitations, the "elan grounding" is a set of wired-in actions.

One of the major goals of this research is to determine these groundings. The process of bootstrapping the implementation of the theory will bring into focus the problems which the base cases must solve.

4. Implementation

4.1 The Importance of the Implementation

The ultimate reason to present an implementation is to show that the recursive theory of rationality can be grounded. This would demonstrate that the theory defines *something*. To the degree that what is thus defined is related to rationality, the implementation would support the viability of the intellectualist approach to rationality. The implementation will contribute to this by demonstrating that a program can realize states and effect state transitions corresponding to those involved in the generation of a rational action according to the intellectualist account I have presented.

For example, a program illustrating the theory should be able to begin in a state corresponding to having a particular intention. It should then enter a state corresponding to deliberating about how to realize that intention. Perhaps it would then deliberate about how to deliberate and reach some possibly wired-in solutions. Thus it performs the deliberations and then performs the actions.

Building an implementation also provides a natural way to bound the research herein proposed. Understanding rationality involves working out or being able to work out the rational act for any rational agent in any situation. This is certainly a monumental task -- each instant every rational agent finds itself in new situations to which it must respond. But those situations and associated rational actions involved in the implementation of a rational agent provide an initial core.

Another reason to implement the theory is to demonstrate its usefulness to more practical artificial intelligence concerns. If the theory really captures some idea of "rationality", then it ought to be helpful in designing programs which are supposed to exhibit intelligence. The motivation for this work came from the supposed improvement in the abilities of systems which would come with the programs being able to reason about their own operations. The implementation of the theory of rationality ought to make clear how this improvement can be achieved.

I must point out however, that the implementation will be a set of particular solutions to more general problems. Many issues will be sidestepped or treated with less rigor than they deserve. Naturally, differences on the importance of the issues thus treated will lead to differences on the relevance of the implementation to the theory or rationality, or of the relevance of the theory itself. I will not present any formal criteria which the implementation must satisfy to somehow "prove" the theory. This implementation is above all an illustration of the theory.

4.2 An Imperative Interpreter

In this section I will present a very rough outline of the program I am constructing to illustrate the theory of rational action. As I have stressed, the implementation is an important part of the research I am proposing. But I am, after all, just *proposing* it -- the program is far from finished. This section is therefore very speculative, very incomplete and may very well be simply wrong in that the ultimate implementation could have little resemblance to that discussed here.

The program under development is called PRAXIS, the term Aristotle used for purposive, rational action. On one hand, PRAXIS will be an "expert" program whose domain is the theory of rational action. It will be an expert in that domain in the same way that other artificial intelligence systems are experts in other domains, say geology or medical diagnosis. So one way to view PRAXIS is as a software illustration of the structural theory -- the user would be able to inspect representations of various states a rational agent would enter in the course of determining the rational action for a situation. The program would be able to represent the relation between an action and its associated deliberations, the conditions for success of an action, the fact of an action being wired-in to a situation, the explicit application of the theory of rationality and so on. Some of these states will be illustrated below.

But in PRAXIS the reasoning *about* action will be linked to the actions of the program. PRAXIS is best described as an "imperative interpreter" because its operations are controlled by a set of rational imperatives which describe what it ought to do. All of the operations of PRAXIS -- "interpreting" the imperatives, deciding what to do based on those imperatives, and then doing what is decided -- are all themselves chosen according to the set of imperatives. Thus PRAXIS is "reflective" in the sense defined by [Smith] in that the interpreter has full access to representations of all aspects of its program and its state and can manipulate those representations to affect its own behavior.

I don't expect PRAXIS to be any better at basic domain-level planning than any existing planning program. It will use the same backward chaining method for planning invented by Aristotle and used by virtually all AI planners. The power of PRAXIS will spring from its flexibility. Most actions of a rational agent are not precisely planned, and even planned actions must give way to improvised responses to unforeseen details of the environment. A rational agent must be prepared to deal with the unexpected. PRAXIS will be so prepared by having access to a richly structured representation of its actions and their associated mental states. With this available the program will be able to notice,

for example, that one of its main goals has been satisfied, even while failing on some subgoal. Rationality is as much a matter of exploiting luck as of careful planning.

PRAXIS will exercise its skills in the domain of automobile travel. The program will "drive" a simulated car through a network of highways. The program will attempt to get from place to place in a reasonable time. The program will have a map of the area but the map may be faulty: it may depict roads which do not exist, or the roads it shows may go to different places. The world is not stable: bridges may be out, roads may be under repair, traffic conditions may make certain routes undesirable. And time constraints (the gas may be running out) will in some cases force the program to decide quickly rather than deliberating or exploring for better solutions. This domain was inspired by the work of [Hayes] which also had an agent traveling in the world and representing its own reasoning process. Part of the appeal of the domain is that it captures the *situatedness* of an agent in the world. Also it is hoped that the metaphor between the movement of an agent through physical space and the progress of a problem-solving agent towards its goal (recognized in the early-AI notion of a "problem space") will facilitate the discovery of analogical situations in the base domain and the higher level deliberative domain. Such situations could be used as the basis of precedential reasoning to improve performance in both domains.

For the moment we will assume that PRAXIS is always in a single well-defined state corresponding to a particular situation. The situation may involve some aspect of the world, or it may correspond to deliberating about some lower-level action. Each rational imperative to which PRAXIS has access has two parts, a relevant *situation* in which the imperative applies, and the corresponding rational *action* to take in that situation.

The following sections describe various aspects of the states PRAXIS can realize and how they are related. This will illustrate various aspects of rational action which the structural theory must be able to represent. I will also present some of the deontic and elan groundings necessary for the system to be given initially so that it will begin running.

4.2.1 The Rational Nexus

Figure 1 illustrates the "rational nexus": a situation S and its corresponding rational action, A. The situation and action are not further described. Thus the structure of situations and actions surrounding the rational nexus surrounds *all* actions.

We represent the state of the agent when it is in situation S as S'. The PRAXIS system will manipulate data structures corresponding to the various states the agent will enter. The data structures could simply consist of a partial list of statements true in the situation. More efficient representations of the state would involving representing states as related to other states in various ways, perhaps as a partial copy or as a member of a hierarchical structure. I don't mean to suggest that the states of rational agents necessarily consist of lists of statements about their situations, however this is how PRAXIS will be implemented.

S' contains information about the situation S, and it is this information which will be emphasized in the discussion below. But S' contains other information as well: It contains information about the currently active intentions of the agent, for example, and perhaps it contains information about recent states the agent has been in. I will individuate the states I discuss below according to the aspect of the state I am emphasizing. But all states will in general contain more information than I explicitly mention.

S is a situation "in the world" and A is an action "in the world", while S' is considered to be

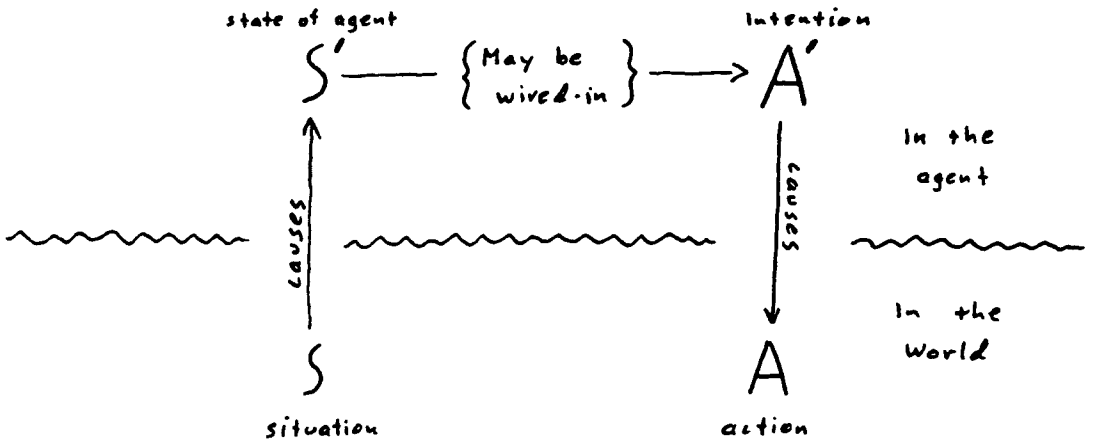

Figure 1.

"internal" to the agent. This separation is not meant to imply any sort of metaphysical dualism, rather it is to keep the representation relations straight. The agent has no access to the situation as such directly. The state of the agent is affected by the situation somehow, it is only those aspects of the situation which have thus affected its state which can have any effect on the action the agent will take.

State A' is an *intention* -- a state of the agent whose conditions of satisfaction include that it causes the action A. For some intentions, the action is as good as taken when the intention forms, other intentions aren't as reliably linked to their actions, and in fact, there might need to be work done by the agent in order to determine how to realize them.

To wire in an action in some situation, the agent (or the programmer) must first determine the state the program will enter when it is in that situation. All potential situations have a wired-in action. In most cases, the wired-in action will be to begin deliberating about the situation as described below. But there will be states where the agent will discover more specific responses to wire-in.

4.2.2 Deliberation

Figure 2 illustrates how deliberative actions are related to the rational nexus. State S', as I have said, may contain representations of the situation. We can speak of the agent as "being in the situation of being in state S'." This description is certainly true whenever the agent is in state S', it might be *useful* to so think of the agent when it must explicitly deal with the fact that it is in that particular situation. In particular when it must *deliberate* about what to do in S, it must manipulate representations of S, these representations are part of state S'. I will (somewhat confusingly, I admit) call both the *state* and the *situation* of being in that state by the same symbol: S'.

State S'' is the state of the agent corresponding to its being in situation S'. Recall that *any* situation in which the agent can find itself can be treated as an instance of a rational nexus. So state S'' stands in the same relation to situation S' as state S' stands to S. In this case, deliberating is a rational thing to do (we'll see why below) and so the rational act associated with S' is that of deliberating about S. Thus perhaps the action of forming an intention to deliberate about S is wired-in to state S''. D(S) represents the state of the agent as it begins to deliberate about what to do in situation S.

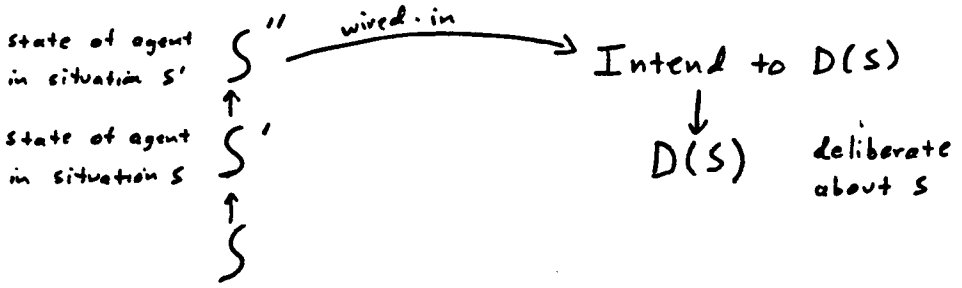

Figure 2.

4.2.3 Application of the Theory

In figure 3, we see the initial deliberation method the program must be supplied with. `D(S)` is wired into a state corresponding to trying to find a representation of an action A such that `R(S) = A`, where `R` is the explicit theory of rationality. When this has been accomplished, the agent must then form an intention to perform A.

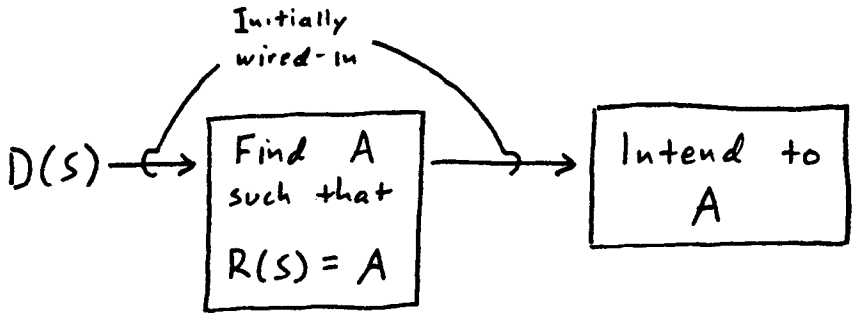

Figure 3.

As I discussed in the previous chapter, applying the explicit theory is in general very difficult. However there are many specific situations in which the agent can conclude (or be told) that a specific action is rational. For example, the agent must be given that the rational act corresponding to state `D(S)` is to discover `R(S)`. To put it quasi-formally:

```
R[D(S)] = find R(S)
```

This particular solution to the problem of applying the explicit theory is an example of the recursive nature of the theory.

Another example of a specific instance of R(S) crucial to the implementation is the action of forming an intention to A, given that R(S) = A and S is the current situation. Indeed the entire operation of the interpreter is controlled by representing each of its steps as specific instances of R(S) for particular situations.

Every time the agent performs an action in a situation S, it has an instance of R(S) for any situation which is represented the same way as S was. This is the rationale behind precedential reasoning. Every specific case of action is stored by PRAXIS as a new rational imperative which it can later apply.

4.2.4 Why Deliberating is Rational

If the state of deliberating about a situation were not wired-in to that situation, the program can still determine that deliberation is rational by the set of actions illustrated in figure 4.

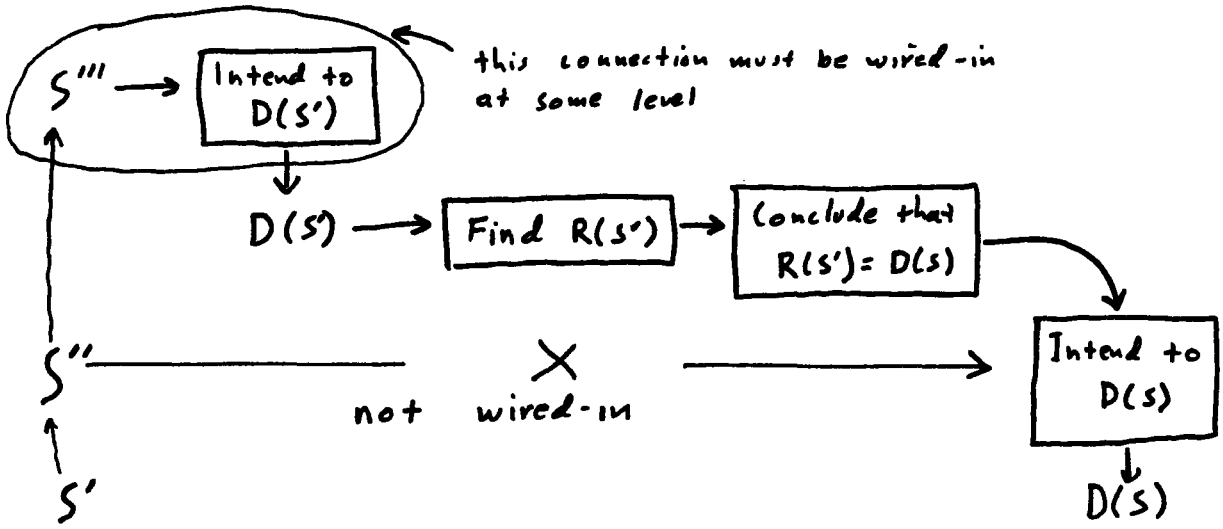

Figure 4.

In this case, assume that state S'' is not wired into an intention to deliberate about S. So now the agent must determine what to do in situation S'. Treating the situation as a rational nexus, we can see that one possibility is to deliberate about S'. The agent can justify that deliberation is rational

based on one of two lines of argument:

1. The agent may have only one thing to do if it does not have a wired-in response, namely to deliberate. So the imperative follows from the agent's limited abilities.

2. The agent may have other things it can do, but deliberating worked in the past -- in other situations in which it had no wired-in direct response. The imperative follows from precedential reasoning.

Either of these arguments could also be the justification for wiring in the deliberation action.

Notice that at some level, for some situations, deliberation *must* be wired-in for the elan grounding of the theory to be accomplished. But it can always be represented *as if* the agent did the wiring-in as a perfectly justified rational action.

4.2.5 Choosing an Imperative

PRAXIS works by interpreting imperatives. Each of them is an instance of R(S) for some particular situation. The imperatives have the force that any rational agent in a situation corresponding to the imperative ought to take the indicated action.

A rational imperative is said to *match* a representation of the current situation if the representation of the current situation entails the situation part of the imperative. So if the current situation is (partially) represented as "It is raining" and an imperative is of the form: "If it is raining or snowing, wear your galoshes!" then the imperative matches the situation. Situations could be represented by lists of logical statements true therein. In this case matching would involve testing that the statements in the situation part of the imperative are logical consequences of the statements in the representation of the current situation.

The method illustrated in figure 5 will be wired-in to the state corresponding to the situation `find R(S)` described above.

First the agent will attempt to match all rational imperative against the representation of S. After this action is complete the agent will be in one of three situations, each of which is associated with an action:

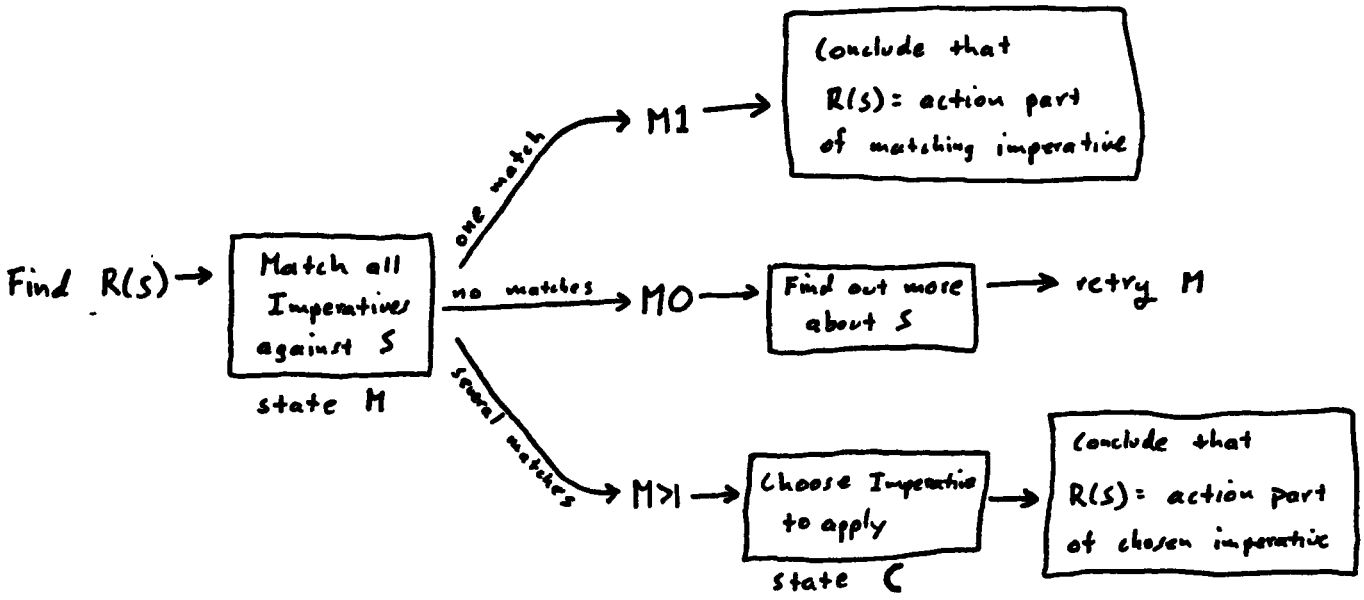

Figure 5.

M1 A single imperative matches S. In this case the agent ought to form an intention to perform the action specified by that imperative.

M0 No imperatives match S. In this case the agent ought to perform actions which will give it more information about S and then attempt to try the matching step again.

M>1 More than one imperative matches S. In this case the agent ought to enter state C, and choose one of the imperatives which match. The result of the choice will be the result of R(S).

In case M1, the forming of an intention to perform the action part of the single matching imperative can be wired-in to the situation. In the other two cases, the formation of the intentions to continue the deliberations in the specified way can be wired-in to the situations specified by the result of the match. These wired-in actions are some of the actions which must be wired-in to provide the elan grounding for the recursion.

A useful imperative which probably must also appear in the initial arsenal of the interpreter is the following:

US If in state C and precisely two imperatives match, choose the more specific one.

Imperative US is consistent with the defeasibility of judgments of rationality -- more information about a situation can change judgments about what it rational.

The action specified above of matching all imperatives against the current situation is certainly difficult given that the agent will have very many such imperatives. I don't mean to suggest that this matching is automatic. The matching process will itself be an action of the agent. And in fact the agent can spend its spare time constructing a more efficient discrimination network to make this matching go as fast as possible. It can determine, for example, a set of features of situations which distinguish large classes of situations, and then wire-in procedures to narrow down the set of possibilities within those classes.

4.2.6 An Example: Overruling a General Imperative

To illustrate how PRAXIS will operate, I will consider a case where a new rational imperative has been adopted:

SP: If in state C and imperatives I186 and I187 match, choose I186.

Note that this imperative applies to the same state, C, as US. It specifies a particular case which the agent may have to consider in state C and tells what to do there. The point of this example is that we assume that I187 is the more specific imperative. The agent may or may not be aware of this and the fact that imperative SP thus contradicts imperative US. As will be seen, the existence of imperative SP will result in the agent never even considering which of I186 or I187 is the more specific.

Now let us consider what will happen when the agent finds itself in a situation in which which both I186 and I187 apply and then begins deliberating. It then enters state C and begins to deliberate about what to do next. The actions which follow are illustrated in figure 6.

Deliberating about C, the program will enter a state C' corresponding to choosing which imperative to apply when reasoning about C. Two imperatives apply: US applies because S has two imperatives (I186 and I187) which apply. But SP also applies to S because it specifically mentions the two rules which apply. (I'll assume that no other imperatives apply.)

More than one rule applies at C', so the agent must begin deliberating about what to do there. In the course of deliberating about C', it enters state C''. In this case, only one imperative will apply to its

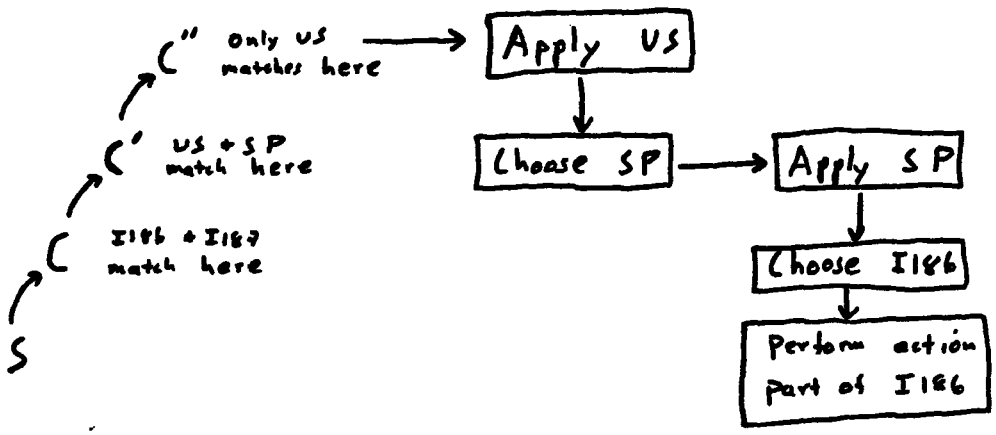

Figure 6.

---

deliberations about C' -- namely imperative US because the imperatives which both apply to C' are US and SP, not the imperatives mentioned by SP. Since only one imperative matches, the system will immediately apply it.

Applying US in this case means that the system must choose the more specific of the two imperatives which match in C' -- the imperatives SP and US. It begins to consider which of the two is the more specific. We assume that it is possible for it to conclude that SP is the more specific (because US applies in all cases where SP does but not vice-versa) and so C' results in SP being chosen. Now that SP has been chosen in C', it will be applied and I186 will be chosen in state S. So the agent will perform the action specified by imperative I186.

This example shows how the explicit representation of the actions of the program allows the same imperative, in this case US, to be to be overruled by another imperative as a result of being *obeyed* in a different situation.

5. Research Proposal

I propose a course of research:

1. I will finish the development and explication of the theoretical and conceptual issues surrounding the notion of rationality and my proposed theory.

Many of the issues dealt with here are philosophical. I believe their discussion is a contribution in itself to some of the philosophical questions surrounding the nature of rationality -- such as action, freedom, reasoning, representation and finiteness. I consider the demonstration that an intellectualist account of rationality can be made to work to be the main contribution of this research.

2. I will develop a model for occurrences in which plans, goals, agents, and the theory of rationality itself can be expressed. I will implement a representational system based on this model of occurrences.

Representational issues have lately become very central to AI research. One of the problems in the methods by which this research is carried out is its typical aim for generality and consequent lack of focus. Simply put: such research is meant to develop systems good for representing *everything*, they are therefore really good at representing very little. My work will emphasize the representation of actions and the processes which go into the generation of actions.

3. I will express the structural theory of rationality in this representational system.

This will involve filling out the details of the situations and actions involved in the generation of rational actions, the explicit application of the theory of rationality, precedential reasoning, and manipulating the computational embodiment of the agent in order to wire-in actions.

Interest in "meta" reasoning is strong in AI, but is almost completely unjustified by the performance of any program which uses it. One of the problems has been the lack of a careful explication of what, exactly, meta reasoning is and how it is related to other kinds of reasoning. This is precisely the job for the structural theory. In it, situations and their associated actions are described and the relations among the actions are set forth.

4. I will experiment in hopes of determining which problems must be solved, and which solutions must be wired-in in order for an agent to be able to ever be able to take an action.

Although my approach to this problem will be empirical, there might be relevant mathematical techniques. Semantics of programs which have access to their own state and context are being developed. [Barwise & Smith, personal communication] An important question any such semantics must answer is if a particular program will ever do *anything*. Perhaps the recursive definition of rationality can also be formalized.

5. I will implement a simulation of a rational agent whose operations are as the theory of rationality describes.

The program will control a representation of the agent. The agent will, as described above, have access to the program which implements it. And the agent will make use of an explicit theory of rationality in order to determine what it will do. The simulation will illustrate the use of precedential reasoning, reasoning about limitations, and wiring-in (and unwiring) actions.

6. Related Work

The work closest in spirit to that of the present proposal is that of Jon Doyle. [Doyle, 1980] presents a theory of introspective rationality covering many of the same issues as I intend to. But Doyle's theory was never implemented, and the failure to implement the theory resulted in several important issues never being addressed. For example, Doyle never demonstrates that his methods for deliberation would ever terminate. He presents a few heuristics for deliberation but gives no indication that they are enough or necessary or even valid. The work I am proposing takes Doyle's work as inspiration and foundation, with a view towards overcoming its limitations.

In [Smith] we find some suggestions about how a fully introspective system could be implemented. Smith's system is a LISP-like interpreter, not a problem solver, but the techniques he uses to implement "reflection" might be relevant to an introspective problem solver. Smith's trick is based on a careful examination of his reflective interpreter to identify the possible states the system may be in when it reflects and gains access to its own state and context. The program is therefore able to construct the state that the "higher level" interpreter would have been in had it been really operating all along.

Fuller discussions of Smith and Doyle, as well as more elaboration of the motivation for introspective problem solvers are found in [Batali, 1983].

Very little of the AI planning literature is relevant to this work. This is because the research has involved very limited models of actions and situations. Research in reasoning about programs has at least involved action models powerful enough to express computational entities, but is still limited. Neither approach has addressed the fundamental problems of determining what is relevant to a problem or how to formalize a problem domain. And planning programs have usually been separated from the performance of the plans they construct. An exception are "robot" languages like AML [Taylor, et al] in which plans are made and then executed, and their progress is tested and compared with expectations.

An area in which recent planning work is relevant is in the modeling of time for planning. [Allen] presents a system in which events occur during and between "intervals" which are partially ordered in time. Similar representations of time have appeared in the philosophical literature (for example: [Rescher & Urquhart]).

[Davis] and [Genesereth] present systems in which "meta" reasoning is done to control and direct the main operations of a program. Their systems illustrate the power which even simple introspective abilities can bring. However in both approaches the meta reasoning is kept separate from the reasoning done at the domain level. Reasoning is not treated as a kind of action, at least not the same kind of action as that done at the domain level.

A number of efforts in AI have viewed analogy as a matter of mappings between representational structures. (For example: [Winston] and [Gentner]) Work on dependency maintenance systems [Doyle, 1979] has involved programs keeping records of the logical dependencies among their inferences. Such dependencies allow the programs to quickly draw inferences when a set of premises is the same as one previously encountered. [Batali, forthcoming] presents a theorem proving system which keeps records of its accomplishments and attempts to prove new theorems by finding analogy-like mappings to previously proved ones.

The main area in artificial intelligence in which sophisticated models of action and Intentionality is to be found is that of natural language understanding. Researchers in this area understood early that little sense could be made of a sentence or story without a decent model of the human world of goals, purposes, and Intentional action. [Wilensky] presents the basis for a natural language understanding system which involves explicit representation of the plans and goals of agents, as well as their "meta" reasoning about those plans and goals.

The theory of action has been a topic for philosophers from the very first. Usually discussed in the domain of ethics, an important problem is the relation of an action to an intention. These and other problems relevant to filling out what I call the "structural theory" of rationality are discussed in [White]. [Searle] presents a version of the "volitional" account of action which holds that an action is a bodily movement caused by an Intentional state. The theory of action I employ is based on Searle's. In fact, the first part of filling out the structural theory will be to attempt to represent Searle's theory of action. The problems involved in practical reasoning, some in what I have called the "structural theory" and a few observations about the "deontic theory" have been receiving attention (see [Raz]).

Existentialist philosophers have emphasized the "finitude" of humans and the attendant problems this raises (see [Barrett]). It is not clear to me if they take note of the fact that an understanding of one's limitations can be used to help decide what to do. [Simon] argues that limitations in available information and information handling abilities should figure in a theory of rational decision making.

[Cherniak] considers how computability and computational complexity issues present limitations for any finite agent. He suggests that a theory of rationality ought to take these considerations into account.

Kant (in [Liddell]) attempts to characterize properties of a rational agent which can be discovered by appeal to the notion of rationality alone, properties which Kant calls *a priori* because they don't depend on particular situations or facts about human biology. The famous categorical imperative has a strong relation to the justification of precedential reasoning. Kant is also an intellectualist and can be read as suggesting that some degree of introspection is involved in rationality.

Searle and Dreyfus, as well as writers in the Existentialist and Pragmatic traditions (see [Bernstein]), emphasize the importance of viewing an agent as embodied in the world, as participating in the causal structure of the world. For such workers *action* is the dominant concern. Far too much of Cognitive Science has inherited the empiricist epistemologist's worry about the validity of knowledge gained through the senses. But it must never be forgotten that the rational agent is immersed in a world to which it must respond. As Karl Marx suggests, the rational agent must do more than "*interpret* the world in various ways; the point is to *change* it."

7. References

James F. Allen, "An Interval-Based Representation of Temporal Knowledge," *Proceedings of IJCAI-81*, 1981.

William Barrett, *Irrational Man*, Doubleday, 1958.

John Batali, *Computational Introspection*, MIT AI Memo 701, 1983.

John Batali, *Theorem Remembering*, forthcoming.

Richard Bernstein, *Praxis and Action*, University of Pennsylvania Press, 1971.

Christopher Cherniak, "Computational Complexity and the Universal Acceptance of Logic," *Journal of Philosophy*, 81(12), December 1984.

R. Davis, "Meta rules: reasoning about control," *Artificial Intelligence* 15, 1980.

Jon Doyle, "A truth maintenance system," *Artificial Intelligence* 12, 1979.

Jon Doyle, *A Model for Deliberation, Action, and Introspection*, MIT AI Technical Report 581, 1980.

Hubert Dreyfus, *What Computers Can't Do, (Revised Edition)*, Harper & Row, 1979.

Dedre Gentner, "Structure Mapping: A Theoretical Framework for Analogy," *Cognitive Science* 7, 1983.

Michael R. Genesereth, "An Overview of Meta-Level Architecture," *Proceedings of AAAI-83*, 1983.

Phillip Hayes, "A Representation for Robot Plans," *Proceedings of the Fourth IJCAI*, 1975.

Brendan Liddell, *Kant on the Foundation of Morality*, Indiana University Press, 1970

Joseph Raz, *Practical Reasoning*, Oxford University Press, 1978.

N. Rescher & A. Urquhart, *Temporal Logic*, Springer-Verlag, 1971.

Gilbert Ryle, *The Concept of Mind*, Hutchinson, 1949.

John Searle, *Intentionality*, Cambridge University Press, 1983.

Herbert A. Simon, "Cognitive Limits on Rationality," in James G. March & Herbert A. Simon, *Organizations*, 1958.

Brian Smith, *Reflection and Semantics in a Procedural Language*, MIT LCS Technical Report 272, 1982.

R. H. Taylor, P. D. Summers, and J. M. Meyer, "AML: A Manufacturing Language," *Robotics*

Research, 1(3), 1982.

Alan R. White, *The Philosophy of Action*, Oxford University Press, 1968.

Robert Wilensky, "Meta-Planning: Representing and Using Knowledge about Planning in Problem Solving and Natural Language Understanding," *Cognitive Science* 5, 1981.

Patrick H. Winston, *Learning and Reasoning by Analogy: The Details*, MIT AI Memo 520, 1979.