Artificial Intelligence and Rational Self-Government

Jon Doyle
March 1988
CMU-CS-88-124

Revised from version of July 6, 1987

Abstract: Using notions from logic and economics, we formulate and formalize the notion of rational self-government, in which the agent reflects on its circumstances, abilities, and limitations to rationally guide its own reasoning and internal organization, in addition to its external actions.

This research was supported by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 4976, Amendment 20, monitored by the Air Force Avionics Laboratory under Contract F33615-87-C-1499. The views and conclusions contained in this document are those of the author, and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Government of the United States of America.

In memory of
Cloe Esther Doyle
teacher, traveler, Tante
Jan. 9, 1907—Dec. 6, 1986

Contents

Preface *v*

1 Introduction **1**

2 Rational self-government **3**
2.1 Mental attitudes 3
2.1.1 Propositions 5
2.1.2 Meanings 6
2.1.3 Consistency 9
2.1.4 Completeness 10
2.2 Reasoning 11
2.2.1 Logic 11
2.2.2 Rationality 12
2.2.3 Attention 15
2.3 Volition 18
2.3.1 Deliberate action 19
2.3.2 Wanton action 20
2.3.3 Rational volition 20
2.3.4 Intentionality 21
2.4 Action and accommodation 22
2.4.1 Conservative accommodation 23
2.4.2 Rational accommodation 25
2.5 Reflection 26
2.5.1 Assumptions 30
2.5.2 Learning 33
2.5.3 Planning 36
2.5.4 Deliberation 38
2.5.5 Introspection 39

3 Constitutional self-government 41
3.1 Laws of thought and legal states 42
3.1.1 Constitutive intentions 43
3.1.2 Satisfying states 43
3.1.3 Conditional attitudes 44
3.2 Constitutive logics of mental states 46
3.2.1 Information systems 46
3.2.2 Satisfaction systems 49
3.3 Laws of thought and legal actions 51
3.3.1 Constitutive priorities 51
3.3.2 Constitutive preferences 52
3.3.3 Conditional actions 53

4 Representative self-government 54
4.1 Social agents 55
4.1.1 Associations 57
4.1.2 Attitudes 58
4.2 Representation 60
4.2.1 Rational representation 62
4.2.2 Self-representation 63

5 Comparative self-government 66
5.1 States of knowledge 68
5.2 Degrees of rationality 69
5.3 Strength of constitution 71

6 Conscious self-government 72

A Immediate implication and inconsistency 75

B Temporal and logical nonmonotonicity 77

References 79

Glossary of some key terms 93

Preface

> The real problem is not whether machines think but whether people do.
>
> Peter Scott, CEO Emhart Corp.

The formal theory of ideal rationality plays a central conceptual role in many mental and social sciences, most vividly in modern politics and economics, but increasingly in psychology, sociology, jurisprudence, and management as well. Its influence is still growing even though its direct application as a theory of human behavior is widely acknowledged to be highly unrealistic. At seeming variance with the ideal theory, human thought is at times habitual, at times rational, with changes of habit requiring effort over time, and with the change-effecting effort available limited by mental resources of memory, attention, and powers of calculation. More realistic theories draw on these criticisms to postulate notions of *limited* rationality, but these suggestions have resisted satisfactory mathematical formalization.

In view of the wide influence of the concept of rationality, it is curious that the field of artificial intelligence, which has as one of its principal aims construction of agents of limited reasoning powers, gives the notion of rationality little explicit role in formulating its ideas, even though these ideas provide relatively precise, if special, settings for understanding limits on rationality. Indeed, one of the major discoveries of artificial intelligence to date has been an appreciation of the power of purely habitual behavior, in the form of carefully tailored but fixed sets of rigid rules, for performing subtle reasoning activities previously thought to be prime examples of ratiocination. Ethologists have long known how apparently complex social and constructive behavior of animals results from small sets of rigid habits, and today's expert systems extend this understanding to many sorts of learned behaviors in humans. But habit alone is not sufficient for understanding the techniques of artificial intelligence. Some techniques,

even some habits, are more simply understood in terms of rational, not habitual, action.

This paper aims to say something both about artificial intelligence and about limited rationality, first by using the concept of rationality to reformulate and clarify several ideas important in artificial intelligence, and by then using these ideas to present initial mathematical formalizations of some aspects of limited rationality in terms as precise as those available for ideal rationality. It is hoped that these ideas will permit further progress on understanding the nature of rationality, as well as offering a basis for construction of more intelligible and reliable artificial agents. In considering these reformulations and characterizations of artificial intelligence ideas, one should keep in mind that current systems, having been developed informally, rarely exhibit the transparency of the concepts presented here. Their opacity is not necessary, and future systems might be designed to exactly capture these structures of reasoning and action.

The central contribution of the following is to articulate, motivate, and formalize the concept of *rational self-government*, in which the agent reflects on its circumstances, abilities, and limitations to rationally guide its own reasoning and internal organization, in addition to its external actions. Rational self-government includes both direct rational control of the path and amount of reasoning performed and indirect guidance through the rational selection of self-regulating "laws of thought" which determine the character and extent of the automatic reasoning performed. As the subsequent discussion indicates, elements of these ideas appear repeatedly in artificial intelligence, psychology, economics, and computer science. (See especially [Baron 1985], [Becker 1976], [March and Simon 1958], [Sternberg 1986], and [Thaler and Shefrin 1981].) The present treatment of these ideas has its origins in my earlier works on introspection, self-regulation, and reasoned assumptions ([Doyle 1979, 1980, 1982, 1983a, 1983d, 1985] and [de Kleer et al. 1977]), with many important ideas drawn from [Minsky 1965, 1975, 1986], [Sussman 1975], and [McDermott 1978]. Further treatments of some of the topics discussed here can be found in [Doyle 1988a,b,c,d,e].

I have attempted to convey these ideas without relying on special expertise in artificial intelligence, psychology, philosophy, logic, decision theory, or computer science, but acquaintance with one or more of these fields will be definitely valuable. For introductions to these fields, readers might consult [Charniak and McDermott 1985] and [Genesereth and Nilsson 1987] on artificial intelligence; [Gazzaniga 1985], [Minsky 1986], and [Hirschman 1982] on psychology; [Har-

man 1986], [Kyburg 1970], and [Searle 1983] on philosophy; [Haack 1978] and [Barwise 1985] on logic; [Luce and Raiffa 1957], [Jeffrey 1983], [Debreu 1959], [Mueller 1979], and [Thurow 1983] on decision theory and economics; [Arrow 1974], [Lindblom 1977], [March and Simon 1958], and [Williamson 1975] on organization theory; and [Harel 1987] and [Garey and Johnson 1979] on computation.

Needless to say, the few concepts studied in this paper, though diverse, do not exhaust the concepts involved in reasoning. In particular, we thoroughly ignore most questions of implementation, among them languages for representing meanings, mechanisms for interpreting these languages, and mechanisms for storing and retrieving elements of memory, including real and virtual representations. These topics, though important, are largely separate from the issues treated here, and some will be treated in subsequent papers. More seriously for the topics of interest, we oversimplify the treatments of propositions, treating them as characterizations of instantaneous states instead of full histories, and do not pursue the question of determinacy or indeterminacy of the agent's states and histories. Both of these latter topics are also to be treated in proper detail in subsequent papers.

In general, there may be an infinite variety of possible reasoners, and though practical experience with a sample of reasoners nourishes the investigator's intuition, only the conceptual, mathematical study of these possibilities can determine which concepts are useful in describing some reasoners, and which concepts, if any, among themselves characterize the full range of possible reasoners. This mathematical study is part of *rational psychology*—the conceptual investigation of psychology—and I hope that this paper, whose organization flows from my own work in rational psychology (beginning with [Doyle 1983a] and [Doyle 1983b]) suggests some of its benefits. (See also [Miller 1986], whose complaints about psychological theories express the problem rational psychology attempts to address.) Virtually every substantial idea discussed in the following appears in either my own work or in the work of others, as cited in the text. But the relative simplicity of the present development—at least as visible when compared to my earlier works—is the deliberate result of the enterprise of rational psychology.

Acknowledgments

I thank Gerald Sussman for past guidance, and Jerry Burch, Jaime Carbonell, Johan de Kleer, Merrick Furst, Clark Glymour, Michael Loui, Ronald Loui, John McCarthy, Drew McDermott, Matthew Mason, Marvin Minsky, Tom Mitchell, B. K. Natarajan, Dana Scott, Peter Szolovits, Richmond Thomason, David Touretzky, and Doug Tygar for many helpful discussions. This paper would not exist but for continuing criticism and encouragement by Joseph Schatz, whose thinking has substantially influenced mine on virtually every topic presented here.

Prior to the circulation of a draft of this paper in July 1987, portions were presented in a class held in January and February of 1987 at Carnegie-Mellon University, and valuable comments and questions were offered by the attendees, especially Larry Eshelman, Oren Etzioni, John Horty, Yumi Iwasaki, Dan Offutt, and Mark Perlin. Gilbert Harman, Steve Minton, Craig Knoblock, and Henrik Nordin offered valuable comments on other presentations of this material.

Chapter 1

Introduction

It is commonplace for people to choose their plans—for the day, for their vacation, for their purchases—on the basis of their tastes. But people choose many other things as well.

- When controversies arise, a person may hear of several explanations or arguments, and may think up some on his own, but in the end settles the question (at least for a while) by choosing the answer he wants to or likes best. Even when the basic facts are agreed upon by all parties, explanations remain controversial because people have different tastes about explanations.

- Many people read about movies (or cars, fashions, foods, celebrities, careers) not just to learn about the movies, but also to learn whether the people they identify with like the movie, and hence whether they should like the movie too. Fears of ostracism and embarrassment can be very powerful, and it is not unusual to find people changing their tastes to conform with their friends.

Thus people's choices do not only concern actions they take in the world, but also how they think and choose as well.

In the following, we explore the notion of rational conduct of reasoning. The standard notion of ideal rationality is that at each instant the agent considers all possible actions, evaluates the likelihood and desirability of each of their consequences, and chooses an action of maximal expected utility. Applied to reasoning, this means that at each instant the agent chooses how it would like

to change its beliefs, preferences, and plans, or change how it represents and formulates things, or how it does things, taking into account the consequences of holding different beliefs, desires, etc. This is the notion of rational self-government.

Often it is not feasible to consider *all* possible actions or all of their consequences, but only some actions, some consequences, and only guesses at their likelihoods and desirabilities. This is the notion of bounded or limited rationality. The amount of time available imposes short term upper limits on rationality (short term because through actions one might be able to gain more time). And how the agent is constituted—the nature of its legal states—provides lower bounds on rationality, the approximation to rationality available without expense of time or attention. Rational self-government modified by constitutional restrictions on possible states is the notion of constitutional self-government.

Bounded rationality can also mean acting with too much information instead of too little. Sometimes the attitudes used to provide guidance come from different sources, and, at least at first, conflict with each other. Until some are changed to remove the conflict, actions can be rational only with respect to subsets of the attitudes. In such cases, the chosen action represents the attitudes the agent chooses to act on. Making actions which represent rationally or constitutionally chosen subsets of attitudes is the notion of representational self-government.

We explore these ideas formally by adapting notions from logic, decision theory, and economics to treat the conduct of reasoning, including some of the ideas and practices of artificial intelligence (reason maintenance, production rules, inheritance links, conceptual hierarchies, constraints, and interpreters). A complete treatment is not possible here, but even the basic elements of rational self-government should permit clearer explications of artificial intelligence ideas and practices than do the usual expositions.

Chapter 2

Rational self-government

In this chapter we set out the basic elements of rational and intentional reasoning and action, along with the basic applications of rational self-government in the mental activities of learning, decision-making, and planning.

To begin with, we view each possible history of the agent as a discrete sequence

$$\ldots, S_{t-1}, S_t, S_{t+1}, \ldots$$

of internal states. We write the set of all possible instantaneous states of the agent as $\mathcal{I}$, and label successive states with successive integers, so that histories are maps from segments of the integers into the state-space $\mathcal{I}$. We need not assume that agents are deterministic. Nondeterministic agents will have several possible histories, agreeing at some instants and diverging at others.

2.1 Mental attitudes

We think of the agent as possessing various mental attitudes at each instant, and view each instantaneous state as a set of attitudes. We will initially consider only three major types of elements: the mental attitudes of *belief*, *desire*, and *intent* (taken in their everyday senses, at least initially). Although there are many other sorts of attitudes (such as hopes, fears, loves, and hates) customarily apparent in human psychologies, not to mention non-attitudinal elements such as emotions and skills, the three attitude types we consider suffice to express the basic structure of rational and intentional action. (See [Searle 1983] and [Harman 1986] for more on these sorts of attitudes.) We divide these three attitude types

into six subtypes, namely the absolute and relative versions of the major types. Absolute beliefs, desires, and intentions are the familiar sorts of propositional attitudes stating that some individual proposition about the world is believed, desired, or intended. Relative beliefs, desires and intentions are comparative attitudes, stating that some propositions are believed more than others, desired more than others, or intended prior to others. We call relative beliefs *likelihoods*; relative desires, *preferences*; and relative intentions, *priorities*.

These six attitude types subsume those of decision theory, action theory, and artificial intelligence. Decision theory (see [Jeffrey 1983]) ordinarily employs only notions of probabilities and utilities, which are numerical representations of likelihoods and preferences. Likelihoods are explicit beliefs about probabilities—specifically, comparative probabilities—and preferences are explicit desires about choices. Action theory (see [Davis 1979]) ordinarily employs only absolute belief and desires, or absolute beliefs, desires, and intentions. Artificial intelligence employs a variety of other names for essentially the same sorts of attitudes. Beliefs are also called facts or statements; desires are also called goals; utilities are called evaluation functions; and intentions are also called plans (both plan steps and plan networks) and agendas. In addition, the elements of programming languages are often anthropomorphically interpreted, data structures as beliefs, imperatives as intentions, conditionals as choices, and procedures as plans. With these identifications, the three major schools of thought in artificial intelligence might be characterized by their concentration on knowledge and reasoning involving pairs of these sets of attitudes. Logicists focus on absolute beliefs and absolute desires; search theorists focus on absolute beliefs and utilities; and proceduralists focus on beliefs and intentions.

Our use of conventional names for elements of mental states is meant to be suggestive, but we employ these names as primitive terms of our theory, not as descriptions. That is, the meanings of these attitudes are given by stipulation of the roles the attitudes play in the psychological organization under discussion. Our central stipulations about the roles these mental attitudes play involve the notions of *volition* or action-taking, and *deliberation* or decision-making.

Some sorts of psychologies involve sorts of attitudes different from those considered here, and our conception of beliefs, desires, and intentions does differ in certain ways from some well-known conceptions. For example, we will consider reasoning in which the agent may change any of its beliefs, desires, or intentions. Such actions are natural in our formal framework, but unconventional to philosophers and economists, who customarily treat belief and preferences as

unchanging or slow-changing properties of agents. We make no assumptions about rates of change, but merely allow that specific attitudes may be easier or harder to change at each instant, and attempt to provide a framework for considering these degrees of fixity.

We let $\mathcal{D}$ stand for the set of all possible mental attitudes, that is, the set of all attitudes we might use in describing the possible states of the agent. Turned around, we view each state $S \in \mathcal{I}$ as a set of mental attitudes, that is, as a subset of $\mathcal{D}$. Thus if $\mathbf{P}\mathcal{D}$ is the powerset (set of all subsets) of $\mathcal{D}$, then $\mathcal{I} \subseteq \mathbf{P}\mathcal{D}$. Let **B**, **D**, and **I** denote the sets (respectively) of all beliefs, desires, and intentions in $\mathcal{D}$. Each of these sets is further divided into subsets of absolute and relative attitudes: $\mathbf{B} = \mathbf{B}^a \cup \mathbf{B}^r$, $\mathbf{D} = \mathbf{D}^a \cup \mathbf{D}^r$, and $\mathbf{I} = \mathbf{I}^a \cup \mathbf{I}^r$, with $\mathbf{B}^a$, $\mathbf{D}^a$, and $\mathbf{I}^a$ respectively plain beliefs, desires, and intentions, and $\mathbf{B}^r$, $\mathbf{D}^r$, and $\mathbf{I}^r$ respectively likelihoods, preferences, and priorities.

2.1.1 Propositions

We view attitudes as attitudes towards propositions, and view propositions as sets of possible worlds, where each possible world decomposes into a state of the agent and a state of its environment. (See [Barwise and Perry 1983] and [Stalnaker 1984] for more on propositions and attitudes.) Formally, to complement the set $\mathcal{I}$ of possible internal states of the agent we let $\mathcal{E}$ be the set of possible states of its environment, and mildly abusing the categories write $\mathcal{W} \subseteq \mathcal{I} \times \mathcal{E}$ for the set of possible worlds. Like $\mathcal{I}$, the sets $\mathcal{E}$ and $\mathcal{W}$ are givens of the theory. Each subset of $\mathcal{W}$ is a proposition, and we write $\mathcal{P} = \mathbf{P}\mathcal{W}$ to mean the set of all propositions.

In a proper treatment we would consider propositions to be sets of possible histories, as these are needed to express expectations of things to come, remembrance of things past, and intended conditions and actions. However, proper formalization of sets of possible histories involves many complexities. For simplicity we restrict this discussion to instantaneous propositions, for these permit us to convey the major ideas even if we slight some important aspects of attitudinal meanings. At such points we indicate that the larger view is needed.

As we will be primarily concerned with the agent's reasoning about itself, the internal portions of propositions will be of more interest than full propositions. We call subsets of $\mathcal{I}$ *internal* propositions, and subsets of $\mathcal{E}$ *external* propositions. If $P \subseteq \mathcal{W}$ is a full proposition, we say that

$$\mathbf{i}(P) = \{S \in \mathcal{I} \mid \exists E \in \mathcal{E} \quad (S, E) \in P\}$$

and

$$\mathbf{e}(P) = \{E \in \mathcal{E} \mid \exists S \in \mathcal{I} \quad (S, E) \in P\}$$

are respectively the internal and external projections of P. We also call these the internal and external propositions determined by P. Propositions purely about the agent's own state satisfy the condition $P = \mathbf{i}(P) \times \mathcal{E}$. Propositions purely about the agent's environment satisfy $P = \mathcal{I} \times \mathbf{e}(P)$. Using these projections, we may write the sets of all internal and external propositions as $\mathbf{i}(\mathcal{P})$ and $\mathbf{e}(\mathcal{P})$ respectively.

We need not make any assumptions in this treatment about the completeness or expressiveness of the universe of attitudes $\mathcal{D}$. That is, we need not assume that every proposition is the content of some attitude, or that every proposition may be expressed through a combination of attitudes. In particular, we will not care whether the agent employs an internal language of thought. Whether or not the agent's attitudes are constructed from such a language, and the structure of the language if they are, are both irrelevant to the issues treated here.

2.1.2 Meanings

We indicate the propositional content of each attitude type by a different meaning function: β, δ, ι to indicate the meanings of (respectively) absolute beliefs, desires, and intentions, and λ, π, and ϖ to indicate the meanings of likelihoods, preferences, and priorities.

The content of an absolute belief (or desire or intention) is a proposition, namely the proposition believed (or desired or intended). We allow meanings of individual attitudes to vary with the agent's state and circumstances, and use three functions β, δ, and ι ot indicate the respective meanings of beliefs, desires, and intentions, so that we have

$$\beta, \delta, \iota : \mathcal{W} \times \mathcal{D} \rightarrow \mathcal{P}.$$

We ordinarily omit reference to the current world, so that if $W = (S, E)$ is the current world, we write $\beta(x)$ instead of $\beta(W, x)$. We allow the possibility that elements of states may encode several different sorts of attitudes. Elements representing single sorts of attitudes are distinguished by having their other interpretations be trivial. For example, if $x \in \mathbf{B}^a$ but $x \notin \mathbf{D}^a$ and $x \notin \mathbf{I}^a$, then we may assume that $\delta(x) = \iota(x) = \mathcal{W}$. If we take into account the division of the world into the agent and its environment, we may define the internal

(external) meanings of an attitude to be the internal (external) projections of its full meanings. That is the internal meanings of $x \in \mathcal{D}$ are just $\mathbf{i}(\beta(x))$, $\mathbf{i}(\delta(x))$, and $\mathbf{i}(\iota(x))$. For simplicity, we will focus on internal meanings, and assume that the internal portions of meanings depend only on the internal state of the agent. Formally, we usually pretend that instead of the earlier definition we have

$$\beta, \delta, \iota : \mathcal{I} \times \mathcal{D} \rightarrow \mathbf{i}(\mathcal{P}).$$

We also note that the propositional content of intentions may be more complex than the contents of beliefs and desires. Intentions may be about either changes the agent intends to effect in its situation or about actions or operations the agent intends to perform. Both sorts of intentions appear in most planned activity, with intended conditions representing the agent's ends, and intended operations representing the agent's means to these ends. In other words, the intended conditions distinguish intentional consequences of actions from unintentional consequences or side-effects. We may think of both aspects combined in single intentions, for example, the intention to open some windows in order to air out the house. In fact, every intention may be at least this complex. As Searle [1983] explains, every intention is self-referential in that in holding it the agent intends that some action be carried out as the result of holding that very intention. Proper treatment of intentions thus requires propositions about possible histories, not just the instantaneous propositions treated here.

Both the contents and treatment of relative attitudes differ from those of absolute attitudes. Relative attitudes compare various propositions, rather than being just about one. More fundamentally, the agent does not perform some action in order to satisfy a preference or priority. Instead, the agent's choice of action to perform either satisfies or frustrates its preferences and priorities. That is, the agent satisfies its relative attitudes through the way it takes actions rather than through which actions it takes.

We assume that each likelihood, preference, and priority specifies a quasi-order (reflexive and transitive relation) over propositions, with worlds in some propositions thought more likely than, more preferred than, or of higher priority than worlds in others. Ordinarily these quasi-orders need not be very complicated, and may, for example, compare only two propositions or each member of two small sets of propositions. We write $\mathcal{Q}$ to mean the set of all quasi-orders over propositions, so that for all $M \subseteq \mathcal{P} \times \mathcal{P}$, $M \in \mathcal{Q}$ iff

1. $(P, P) \in M$ for each $P \in \mathcal{P}$, and

2. For each $P, Q, R \in \mathcal{P}$, if $(P, Q) \in M$ and $(Q, R) \in M$, then $(P, R) \in M$.

Thus if $M \in \mathcal{Q}$, the first condition of the definition says that M is a reflexive relation, and the second condition that M is transitive. We interpret likelihoods, preferences, and priorities via respective meaning functions

$$\lambda, \pi, \varpi : \mathcal{W} \times \mathcal{D} \to \mathcal{Q}$$

where as before we ordinarily elide reference to the agent's current circumstances and write, for example, $\lambda(x)$ instead of $\lambda(W, x)$.

When we have a specific $M \in \mathcal{Q}$ in mind, for ease of reading we often write $P \lesssim Q$ instead of $(P, Q) \in M$, no matter which sort of attitude is being discussed. In general, neither $P \lesssim Q$ nor $Q \lesssim P$ need hold in a quasi-order, but if both $P \lesssim Q$ and $Q \lesssim P$ hold, we write $P \sim Q$ and say that P and Q are *equivalent* in the quasi-order, and if $P \lesssim Q$ holds but not $Q \lesssim P$, we write $P < Q$ and say that Q is *greater* than P in the quasi-order. We partition each $M \in \mathcal{Q}$ into the strictly greater and equivalence relations $M^<$ and $M^\sim$ by defining

$$M^< = \{(P, Q) \in M \mid (Q, P) \notin M\}$$

and

$$M^\sim = \{(P, Q) \in M \mid (Q, P) \in M\}.$$

Clearly, $M = M^< \cup M^\sim$ and $M^< \cap M^\sim = \emptyset$. We decompose λ into two subsidiary meaning functions $\lambda^<$ and $\lambda^\sim$ by defining $\lambda^<(x) = (\lambda(x))^<$ and $\lambda^\sim(x) = (\lambda(x))^\sim$ for each $x \in \mathcal{D}$, and in the same way, we get $\pi^<$ and $\pi^\sim$ from π, and $\varpi^<$ and $\varpi^\sim$ from ϖ.

For convenience in describing orders, we let $1_{\mathcal{Q}} \in \mathcal{Q}$ stand for the identity relation on propositions,

$$1_{\mathcal{Q}} = \{(P, P) \mid P \in \mathcal{P}\}.$$

If $M \in \mathcal{P} \times \mathcal{P}$ is a relation over propositions, we write M^* to mean the transitive closure of M, that is, the least superset of M containing (P, R) whenever it also contains (P, Q) and (Q, R).

For example, an element x interpreted as a preference for sun over rain would have a meaning

$$\pi(x) = \{(P_{\text{rain}}, P_{\text{sun}})\} \cup 1_{\mathcal{Q}}$$

where P_{sun} is the set of all worlds in which the sun is shining and P_{rain} is the set of all worlds in which it is raining. An element x interpreted as indifference or neutrality between pleasing one of one's n children over another would have a meaning

$$\pi(x) = \{(P_i, P_j) \mid 1 \leq i, j \leq n\} \cup 1_Q$$

where for each i, $1 \leq i \leq n$, P_i is the set of worlds in which one's i'th child is happy. This order says that pleasing one child is of equivalent desirability to pleasing any other. (It might appear to some more natural to define $\pi(x) = 1_Q$ to represent indifference, but "not preferred to" is a different notion than "as preferred as.")

2.1.3 Consistency

The theory of ideal rationality imposes various consistency conditions on the attitudes a rational agent may hold simultaneously. (We later will consider agents in which these conditions are not always met.) The basic consistency conditions are that the set of instantaneous absolute beliefs be consistent, and that each of the sets of current likelihoods, preferences, and priorities be individually consistent.

The agent's absolute beliefs in the current state S are consistent just in case some world satisfies them all, that is, just in case the propositions $\beta(x)$ for $x \in S$ have nonempty intersection, that is, if

$$\beta(S) = \bigcap_{x \in S} \beta(x) \neq \emptyset.$$

The agent's likelihoods are consistent just in case the combined strict likelihood relation

$$\lambda^{<}(S) = (\bigcup_{x \in S} \lambda^{<}(x))^*$$

is also a strict likelihood relation and is disjoint from the combined equivalent likelihood relation

$$\lambda^{\sim}(S) = (\bigcup_{x \in S} \lambda^{\sim}(x))^*,$$

that is, $\lambda^{<}(S) \cap \lambda^{\sim}(S) = \emptyset$, so that there are no propositions P, Q such that $P \sim Q$ and $P < Q$ in the combined orders. Consistency of preferences and priorities is defined similarly.

There are other sorts of consistency conditions one might impose as well. For example, the intentions of the agent should be consistent, but since different intentions may refer to different times, this condition is not easy to state without going into the structure of propositions about histories. One might go further and require that the agent's beliefs be consistent with its intentions, in the sense that if one intends something one must believe that it is possible, and might believe that it will occur. Or one might require that one's absolute and relative beliefs be consistent in the sense that one's doubts cannot be likelier than one's beliefs, and that the extremal propositions $\emptyset$ and $\mathcal{W}$ must be judged respectively unlikeliest and likeliest. These conditions and others are discussed in the literature, but we will not pursue them here, since most are not directly relevant to the questions we consider, and some are controversial proposals.

In any event, each of the consistency conditions defined above is a substantive condition, that is, each refers to the meanings of attitudes. In section 3.2 we consider non-substantive or internal consistency conditions on attitudes as more feasible demands on the coherence of the agent. These non-substantive conditions are expressed directly as logics of states without reference to the meanings of attitudes.

2.1.4 Completeness

When the agent's attitudes in state S are consistent, one may find numerical rankings or measures of degrees of likelihood, preferability, and priority compatible with the combined orders $\lambda(S)$, $\pi(S)$, and $\varpi(S)$. These are functions $p, u, q : \mathcal{P} \to \mathbf{R}$ such that $p(P) \leq p(Q)$ whenever $P \lesssim Q$ in $\lambda(S)$, $u(P) \leq u(Q)$ whenever $P \lesssim Q$ in $\pi(S)$, and $q(P) \leq q(Q)$ whenever $P \lesssim Q$ in $\varpi(S)$. Rankings of preferability and priority are called, respectively, *utility* and *priority* functions. Rankings of likelihood are called *degrees of belief*. Subjective *probability* measures are rankings of likelihood in the interval $[0, 1]$ that satisfy the conditions that $p(\emptyset) = 0$, $p(\mathcal{W}) = 1$, and $p(P \cup Q) = p(P) + p(Q)$ when $P \cap Q = \emptyset$. (See [von Neumann and Morgenstern 1944], [Savage 1972], and [Shafer 1976].)

A quasi-order $\lesssim$ is said to be *complete* if for every $P, Q \in \mathcal{P}$ either $P \lesssim Q$ or $Q \lesssim P$ holds. In the straightforward uses of relative attitudes in artificial intelligence, the agent's knowledge about the world is limited, and the orders $\lambda(S)$, $\pi(S)$, and $\varpi(S)$ are rarely complete. In this case, there may be many numerical rankings compatible with each order, yielding sets $\hat{p}(S)$, $\hat{u}(S)$, and

$\hat{q}(S)$ of compatible ranking functions with, for example, p compatible with $\lambda(S)$ for each $p \in \hat{p}(S)$. However, alternative ranking functions may be incompatible with each other on propositions not related in the quasi-order.

Ordinary decision theory makes stronger assumptions about the completeness of the agent's attitudes than artificial intelligence. In particular, the usual starting point in decision theory is a set of likelihoods that is both complete and "fine," which means that the set of likelihoods orders a continuum of possibilities. Under these assumptions it is proved that the set of likelihoods determines a unique probability measure p on propositions. Similarly, standard decision theory assumes the preference relation is complete and fine as well. With this assumption, any utility functions compatible with the partial order of preference are compatible with each other, that is, if $u, u' \in \hat{u}(S)$, then $u(P) \leq u(Q)$ iff $u'(P) \leq u'(Q)$. In this way the strong completeness assumptions of standard decision theory serve as uniqueness assumptions about the agent's comparisons of propositions.

2.2 Reasoning

We call the agent's changes of attitudes from state to state *changes of view*. We assume that while some of these changes may happen to the agent, due to external or bodily influences, some changes are the result or realization of the agent's reasoning. Instant-to-instant changes may correspond to the elementary steps of reasoning, while larger subhistories of the agent may correspond to bigger steps of reasoning. In the following, we will concentrate on reasoning by assuming that each change of attitudes under discussion represents reasoning.

2.2.1 Logic

It is conventional to view logic as a theory of thinking setting out the laws of thought and principles of reasoning. In this view, psychology is actually a branch of logic, with mental objects taken to be sentences in a logical language, mental operations taken to be inferences in a formal logical system, and with reasoning being purely a matter of automated deduction. But this view is misleading, for reasoning involves changes of desires and intentions as well as changes of beliefs, and reasoning is not cumulative but is instead often nonmonotonic.[1]

[1] See Appendix B for a definition and discussion of nonmonotonicity.

(See also [Thomason 1987].) That is, mental attitudes may be abandoned as well as gained. Utterly ordinary nonmonotonic changes of beliefs, for example, occur as the agent accounts for the effects of its actions and the results of its observations. In the latter case, the observational results themselves must be fit into the agent's beliefs, displacing previous beliefs, while in the former case, the agent must update its beliefs to accommodate the expected consequences of its actions, many of which involve changes in the world. Thus reasoning is not a matter of automated deduction, which in the standard conceptions involves only beliefs and which in any event involves only additive changes of attitudes. As Harman [1986] puts it, inference is not implication: reasoning and inference are activities, *reasoned* changes in view, while proofs in a logic are not activities but atemporal structures of a formal system, distinct from the activity of constructing proofs. Thus logic is not, and cannot be, the standard for reasoning.

Logic, of course, may be employed to formalize psychological theories. For example, logics might be formulated to describe the instantaneous closure and consistency properties of or implications of agent's attitudes, such as the consistency conditions on states related to rationality. In such cases the logic characterizes the structure of the agent's states. Even though reasoning is not cumulative, the logic of the agent's individual states need not require the use of any sort of non-standard logic, since nonmonotonic changes in states may occur even when states are closed and consistent with respect to a deductive logic. But this use of logic is not particular to psychology, for in the same way logic may be used to formalize meteorology or any other subject matter, and mental operations are not thereby inherently logical operations any more than meteorological events are thereby inherently logical operations. Even though one may use a logical theory of the agent to deduce properties of its histories, including its next state, this external deduction is not an implementation of the agent. The agent is a model of the theory, not the theory itself, and if the two are confused, the agent must be cumulative, incrementally calculating partial but increasingly complete histories. In such cases there is no way to say that the agent possesses any attitudes at all, since it does not possess a single instantaneous state.

2.2.2 Rationality

The idea that steps of reasoning are not logical operations is disturbing to many people, for if reasoning need not be logical, it could be anything. This is true:

any change of beliefs, for example, might be a step of reasoning. But one need not immediately abandon all standards of soundness when one distinguishes reasoning and logic, for a better standard for reasoning is rationality, which in its simplest form means taking actions of maximal expected utility. That is, we view reasoning as an activity to be governed like other actions by means of the agent's intentions and plans, and we aim for this government to be rational by making the formation and execution of plans rational. In the simplest case, we may ignore the question of intent and think of reasoning as change of view in which the selection of each successor state is rational. In this case, steps of reasoning are considered to be actions which change only the internal state of the agent, and the reasoning actions possible from the current state are compared to find the action of maximal expected utility.

What is lacking in logic as even an ideal theory of thinking is that reasoning has a purpose, and that purpose is not just to draw further conclusions or answer posed questions. The purpose or aim of thinking is to increase insight or understanding, to improve one's view, so that, for instance, answering the questions of interest is easy, not difficult. Rationally guided reasoning constantly seeks better and better ways of thinking, deciding, and acting, discarding old ways and inventing and adopting new ones. This view of reasoning has been argued and examined at length by numerous authors, including Minsky [1975], Hamming [1962], Rorty [1979], and Harman [1973, 1986]. Truesdell [1984, pp. 498-499], for example, emphasizes that reasoning, like scientific inquiry, is a sequence of *rational* guesses and *rational* revisions, not merely a sequence of guesses and revisions, with the bases for rational selections of guesses and revisions found in the agent's experience (personal or vicarious), and with this experience rationally selected and pursued as well. Polya [1965] agrees, devoting an entire volume of his book on mathematical discovery to the subject of rational investigation. (The same view applies more generally to much of human activity in addition to reasoning. See [Schumpeter 1934], [Drucker 1985], [Becker 1976], [Stigler and Becker 1977]; also [de Sousa 1987], [Ellis and Harper 1961], [Peck 1978], [Russell 1930], and [Yates 1985].) Note that rationality in reasoning does not mean that the agent's beliefs about its current state are either accurate or complete, merely that the choices of reasoning steps are rational given these imperfect beliefs.

This conception of reasoning is very different from incremental deduction of implications. Guesses, rational or not, are logically unsound, and instead of preserving truth, reasoning revisions destroy and abandon old ways of thought

to make possible invention and adoption of new ways of thought. For example, agents may waver and waffle, repeatedly changing their minds on questions, yet still make progress with each change. Rational deliberation (see section 2.5.4) illustrates this vividly, for in deliberation the agent constantly reverses its tentative decisions as it ascends the "ladder of wit," as Barth [1967, pp. 238-239] calls it. One might hope to organize reasoning to avoid non-logical assumptions and revisions, but it hardly seems possible to live any other way. Guesses are necessary, for humans at least, because of the frailty and smallness of human mental capacities. Denied complete and certain knowledge we assume our way through life, only dimly and occasionally aware through our meager senses of any reality, and even then loath to part with our cherished beliefs. Revisions, in turn, are necessary because even if guesses are never wrong, progress in reasoning, like maturity and progress in life, requires escape from the shackles of the past. Cumulative agents, being unwilling to give up the past, are condemned to repeat it endlessly. Put most starkly, reasoning aims at increasing our understanding; rules of logic the exact opposite.

As in ordinary decision theory, we say that rational actions are actions of maximal expected utility, where the expected utility of an action is the average of the utilities of its effects discounted by the probabilities of their occurrence. (See [Jeffrey 1983], [Berger 1985].) We think of actions as functions $a : \mathcal{W} \to \mathcal{W}$ from worlds to worlds, or more generally, if we wish to include nondeterministic actions, as correspondences $a : \mathcal{W} \to \mathbf{P}\mathcal{W}$. Each action a is thus a subset of $\mathcal{W} \times \mathcal{W}$. Borrowing from dynamic logic ([Harel 1984]), we discuss the effects of actions by means of expressions like $[a]P$, which is read as "P is true in every state reached by taking a," and $\langle a \rangle P$, which is read as "P is true in some state reached by taking a." $[a]P$ and $\langle a \rangle P$ are both propositions if P is, namely

$$\langle a \rangle P = \{W \in \mathcal{W} \mid \exists W' \in P \quad (W, W') \in a\}$$

and

$$[a]P = \{W \in \mathcal{W} \mid \forall W' \in \mathcal{W} \quad (W, W') \in a \supset W' \in P\}.$$

Given a utility function u and a probability measure p, the expected utility $\bar{u}(a)$ of action a is then just

$$\bar{u}(a) = \sum_{W \in \mathcal{W}} u(\{W\}) \cdot p([a]\{W\}),$$

that is, the utility of the expected consequences of a averaged over all atomic propositions $\{W\}$. But in our setting, each state determines a set $\hat{u}(X)$ of compatible utility functions, and a set $\hat{p}(X)$ of compatible probability distributions.

In this situation we say that a is a rational action if its expected utility is not exceeded by the expected utility of any other a' under any choices of $u \in \hat{u}(X)$ and $p \in \hat{p}(X)$. That is, a is rational in state X if for every other action a'

$$\sum_{W \in \mathcal{W}} u(\{W\}) \cdot p([a]\{W\}) \geq \sum_{W \in \mathcal{W}} u'(\{W\}) \cdot p'([a']\{W\})$$

for every $u' \in \hat{u}(X)$ and $p' \in \hat{p}(X)$. Unfortunately, this is a very strong condition, and in general there may not be any actions rational in a given state. That is, incompleteness of the agent's attitudes leads to nonexistence of rational actions. In ordinary decision theory, the conditions imposed on likelihoods and preferences mean that any ranking function compatible with the orders gives the same result, so there is at least one action rational in each state. In the more general case, the agent may have to make assumptions, adopt temporary likelihoods and preferences, in order to act rationally. In section 2.4.1 we discuss this further as the "will to believe." (See also [Seidenfeld, Kadane, and Schervish 1986] for a treatment of decision theory that permits incompleteness of attitudes.)

2.2.3 Attention

Thus reasoning, even in perfectly rational agents, is a complex affair, and is rarely as simple as mechanical application of the rules of some system of logic. But it also is not simply a mechanical application of the formulas of decision theory in the sense of having agents calculate what is rational in order to decide what to do. The ideal rationality studied by economists has a precise and deep mathematical theory, but is unmechanizable, as it requires too much information on our part to feasibly construct agents, and too much computation on the part of the agent for it to act. Ideal rationality requires the agent to possess enormous numbers of probabilities and conditional probabilities about the propositions involved in reasoning. Many of these are unconsidered and unknown (what is the probability that a gazelle has cancer?), hence difficult to obtain, and too numerous to store in memory using direct methods. Moreover, the calculations involved are too difficult, no matter how big and fast a computer is used. Many reasoning problems demand that the reasoner discover a sequence of inferences in a space of possible paths that grows exponentially with increasing length, so that even huge increases in computer speed and size purchase only modest improvements in reasoning powers.

These difficulties arise whether one aims to rationally choose external actions or internal steps of reasoning. In the case of rational selection of successor states

the problem may appear to be easier, for the actions under consideration are quite definite, yielding specific resultant states. But the consequences of interest may not be immediately visible in individual states of the agent. If the agent desires to solve some intellectual problem, or see if a chess board has a forced win, it may know that its hypotheses are sufficient to determine the answer, but not be able to tell what the answer is. This means that even if the agent can evaluate the utility of any proposition, including the utility of its current state regarded as a proposition $\{S\}$, it may not be able to determine the likelihoods of consequences of this state, and so be unable to determine the expected utility of its possible actions.

Early on Simon [1969, 1982] suggested *bounded* rationality as a more realistic notion, that is, thought and action rational given limited information, memory, computational ability, and time. (See also [Cherniak 1986].) In this view, the exact calculations of expected utilities is replaced by a process of estimating expected utilities called *search*. There is a large literature on sequential decision problems and search processes, which we will not treat or review here: see [Pearl 1984] and [Raiffa 1968].

Unfortunately, reasoning and search are sometimes identified in the literature, with search defined to be choosing what to do next, a definition that can be viewed as subsuming reasoning. But it is better to separate these notions than to conflate them. Almost all concrete discussions of search, and almost all specific applications of search in artificial intelligence, involve search for a solution to a specific goal by examining the elements of a space of possible or partial solutions. One can of course view reasoning as search in this sense, but only by trivializing search to include all motion. In reasoning there need not be a fixed goal toward which reasoning strives. Instead, the agent may change any of its attitudes, including its desires and intentions, at each step. This means that reasoning is not search through a fixed search or problem space, but rather a series of partial searches through a series of different search or problem spaces, potentially reformulating and replacing the search space at each instant, so that no progress is made within any one of them. One could call this search, but reasoning seems a more apt and less presumptuous label.

In the context of decision making, reasoning is the process by which decisions are formulated, while search is a process by which decisions are evaluated (even if only approximately). The same distinction underlies the two fields of decision analysis and decision theory. Infeasibility aside, the theory of ideal rationality is unhelpful in mechanizing action, as it is a theory for evaluating

decisions, not coming to them. The process of constructing an agent, and of the agent reaching a decision, are processes of developing a formulation of the decision, of developing the alternatives, values, and expectations relevant to the current situation. But decision theory presupposes the agent's alternatives, values, and expectations. It is a theory of their import, not a theory of their construction. In contrast, the concern of the field of decision analysis is specifically the process of constructing good formulations of decisions (see [Raiffa 1968] and [Howard 1980]). Indeed, many ideas and techniques in decision analysis strongly resemble some ideas and techniques for reasoning studied in artificial intelligence, especially in the area of knowledge acquisition, but the two fields have until recently developed with little or no communication.

One of the principal means for limiting rationality is to limit attention and effort by directing them to some areas and ignoring others. For example, searching large spaces without direction is quite difficult, for there is no way to set strategy, or choose "islands" ([Minsky 1961]) to guide the search. Without global direction to smooth out these irregularities, search must follow every minute detail of the space. Focusing attention increases efficiency, but at a price, since the path of reasoning may diverge greatly from paths it would have taken had the ignored areas been considered, and as a consequence the agent must occasionally suffer nasty surprises, crises, or comeuppances. Leibenstein [1980] calls such measured attention *selective* rationality, and develops a good bit of economics from its base.

Plans and procedures are the means by which such directions are expressed. The sequences of steps and the structures of tests and branches prescribed by simple procedures are rigid, and ignore many possibilities that rational action or search might consider. Procedural government of reasoning thus corresponds to the notion of attention, in which the agent decides to concentrate its effort on one task to the temporary exclusion of others.

Unfortunately, just as a few logical aspects of thinking encouraged the mistaken view that reasoning is logical, the procedural aspects of thinking have encouraged a view that reasoning must be computational. The result is the common dogma that artificial intelligence seeks to understand thinking as an effective computational process. This view may also be mistaken. The fundamental use of procedures in thinking is to limit attention and so conserve resources. This use does not depend on what sorts of steps or resources are involved, and if the elementary steps or operations are transcomputable, reasoning may be procedural yet not be effectively computable. The theory of computation focuses

on particularly simple steps, namely those comparable to Turing machine operations, and to resources of time and space derived from these steps. But there are many more sorts of steps and resources in the world than Turing machine operations and Turing machine time and space, and there is no compelling reason at present to suppose that all of the elementary sorts of operations available in human reasoning are reducible to Turing machine operations. The common claim of artificial intelligence that thinking is computational is actually irrelevant to the field's present successes.

More generally, it is worth understanding that the infeasibility of ideal rationality and the notion of attention are matters of economics, not matters of computability. Whether or not some elementary operation available to the agent is computable, performing the operation entails some cost in terms of the agent's resources. To conserve its resources, the agent must allocate them wisely, that is, focus its attention on some things and expend no effort on others. Thus the notion of attention arises from the economics of resource allocation even if none of the agent's elementary operations are computable. (See also [Winograd and Flores 1986].)

2.3 Volition

We call the process by which actions are taken in order to satisfy the agent's attitudes *volition*. (See [Searle 1983] and [Dennett 1984].) Different sorts of volition result when actions are taken to satisfy different sorts of attitudes. The two main sorts of action are *deliberate* and *wanton* action, where the distinction between these two ways of acting turns on whether intentions or desires determine the agent's volition.

We discuss these varieties of volition below, but first note that in contrast to the usual assumption in automata theory, we do not think of the actions of the agent as described by a transition function $\tau : \mathcal{I} \rightarrow \mathcal{I}$ such that $S_{t+1} = \tau(S_t)$ if deterministic, or by a transition correspondence $\tau : \mathcal{I} \rightarrow \mathbf{P}\mathcal{I}$ such that $S_{t+1} \in \tau(S_t)$ if nondeterministic. The automaton view of action makes a strong assumption, namely that the agent has no implicit memory or state. That is, transition correspondences imply that the changes possible at an instant depend only on the internal state of the agent at that instant, regardless of the past history of the agent. Mathematically, of course, we may identify behavioral equivalence classes of automaton states with behavioral equivalence classes of

agent histories, but if we only make a portion of the agent's state explicit, then the transitions possible at an instant depend on the past history of the system, so that transition rules are functionals of histories. Our conception of the agent's actions exhibits such history dependence, since the internal meanings of attitudes used in selecting actions may depend upon the external environment, in which case the agent uses the environment to store information. (See also [Rosenschein and Kaelbling 1986].)

2.3.1 Deliberate action

In deliberate action, volition means taking actions to satisfy the agent's absolute intentions in an order that satisfies the agent's priorities. We think of the actions of an agent as resulting from it executing a *volitional procedure* over its current attitudes. An abstract but typical volitional procedure for deliberate action might be as follows.

1. Select the next intention to carry out.
2. Select the method by which to carry out the selected intention.
3. Carry out the selected intention using the selected method.
4. Repeat these steps.

Many variants or refinements of this procedure are possible, each reflecting a different degree of complexity in acting, and different degree of utilization of the agent's powers of reasoning. The minimal model of the volition procedure is that familiar as programming language interpreter (see [Abelson and Sussman 1985]). The central process of such interpreters is to repetitively calculate the next instruction, retrieve its meaning (the procedure's code or body), and then to execute the procedure. In most programming language interpreters, the selection steps are trivial since instructions (intentions) are arranged serially by design, and each has a single method (machine instruction or procedure body) attached. More complex volitional procedures interleave the steps above with steps for formulating and selecting new intentions, priorities, and methods, or otherwise revising the current sets of intentions and priorities. (See, for example, [McDermott 1978], [Doyle 1980], and [Batali 1985].)

2.3.2 Wanton action

In contrast to deliberate action, in wanton action the agent acts on the basis of its desires and beliefs, not on the basis of intentions or plans. A volitional procedure for wanton action might be as follows.

1. Select a desire.
2. Select an action relevant to satisfying the selected desire.
3. Carry out the selected action.
4. Repeat these steps.

As with the volitional procedure for deliberate action, many variations and refinements of this procedure are possible. One nearly minimal model of wanton action is GPS [Newell and Simon 1963], in which relevance of actions (operators) to desires (goals) is stated in a fixed table of "differences." More complex procedures interleave these steps with steps for formulating and selecting new desires and methods, for example, adopting new desires as subgoals for satisfying an existing desire.

2.3.3 Rational volition

Though they provide different bases for action, neither deliberate nor wanton action need be rational, though each may be. Deliberate action need not be rational, for the selections of plan and method might be irrational. Indeed, in some cases people carry out plans without regard to their own judgments of rationality. This sort of behavior is described as compulsive, mechanical, robot-like, or "going through the motions." When desires and intentions are both considered but conflict, as when a dieter is confronted with an especially tempting but forbidden dessert, we speak of the outcome as demonstrating weakness of will when the intentions are not followed and as demonstrating strength of will when they are.

Wanton action may be more or less rational depending on how the desire and action are selected. When the strongest or maximally preferred desire is chosen and the action is selected to be of maximal expected utility towards satisfying the chosen desire, wanton action can be much like rational action in the standard sense. When rationality is limited, the agent possesses no clear or consistent

notion of strength, and only imperfect knowledge about relevant actions, so the actions actually taken may not appear very coherent. At one instant, the agent may act on one desire, while at the next instant, on another, possibly contrary, desire, and the actions taken may only be first steps along a path leading to satisfaction with no guarantee the agent will ever follow the rest of the path.

2.3.4 Intentionality

Of course, wanton action can be viewed as deliberate action in which the agent never forms any intentions not immediately acted upon, and some theories of intentional action include such "intentions in action" (in Searle's [1983] terminology). The major difference between purely rational action and governed action, whether deliberate or wanton, is that in volition the agent distinguishes between the intended changes wrought by some action and all other changes wrought by the action, the unintended changes. This distinction is outside of the standard theory of rational action, where all effects of actions receive equal treatment. In deliberate action, the intended conditions are stated explicitly. Even in wanton action, one may use desires to interpret intentionality of action effects, and some philosophers have proposed theories of action that involve only beliefs and desires, making elaborate constructions attempting to discriminate intentionality of effects purely on the basis of beliefs and desires. See [Goldman 1970] and [Davis 1979].

The problem of intentionality arises in artificial intelligence as well, where most systems lack the distinction between intended and unintended effects. These systems typically represent actions by sets of partial changes or by simple procedures. For example, STRIPS [Fikes and Nilsson 1971] represents actions by "addlists" and "deletelists," lists of statements to be added to and subtracted from the current description of the world. But these lists represent only expected changes, not intended changes. That is, the changes the action is intended to bring about may or may not be explicitly represented by inclusion in the addlist or deletelist, and may not be the only things there. Similarly, HACKER [Sussman 1975] represents actions with LISP procedures that when executed change the state appropriately. But these procedures simply change the state, without distinguishing between unintended changes and intended changes. Indeed, in correcting its errors, HACKER performs elaborate analysis of its goals attempting to reconstruct just which changes were intended and which were side-effects, an analysis that strongly resembles constructions made by the philosophers of

action.

2.4 Action and accommodation

Intended actions are found by interpreting the intention selected by the volitional procedure in the context of the current state. Thus if x is the selected intention, we define the set of possible successor states Γ_t to be

$$\Gamma_t = \{S \in \mathcal{I} \mid (S_t, S) \in \iota(x)\},$$

and require that the next state S_{t+1} be an element of Γ_t. Observe that here we must explicitly recognize that intentions refer to sets of possible histories, not simply to sets of instantaneous possible worlds.

The condition $S_{t+1} \in \Gamma_t$ on state changes may badly underdetermine the revised set S_{t+1}. Most plans, of course, are incomplete as descriptions of the changes that take place, as they stipulate only a few explicit changes that the agent intends to bring about, and leave other changes unstated. For example, if the selected intention is that some belief $y \in \mathbf{B}^a$ should be adopted, then choosing $S_{t+1} = \{y\}$ (giving up all elements not reasserted) satisfies these restrictions on changes, assuming for the sake of argument that $\{y\} \in \mathcal{I}$. It is reasonable to ignore side-effects of intended changes initially because in everyday life most actions make only limited changes in the world, leaving most conditions unchanged. But because the parts of the world are not wholly independent, sometimes intended changes inevitably entail unintended, even undesired changes. Sometimes special mechanisms (such as requiring prescriptions for medication and impact statements for major construction) are imposed to remind people to consider the side-effects of otherwise desirable actions.

Part of the reason for this incompleteness is that it is often very difficult to describe or predict the effects of actions in specific situations. The field of artificial intelligence has expended much effort toward axiomatizing the effects of mundane sorts of actions, and one of the insights gained in this effort is that ordinary knowledge about actions is highly qualified and incomplete. That is, common knowledge about an action describes its usual effects, but with the qualification that nothing untoward or exceptional occurs. The simplest of these descriptions are of the form of "laws of motion," which we may write as formulas $p \supset [a]r$ stating that if p is true in the current state, then after taking action a, r will be true of the agent's future history. Most such expectations are also

qualified, either implicitly or explicitly with one or more qualifying conditions as in $(p \wedge \neg q) \supset [a]r$, but in general not enough is known to determine either that p holds or that q does not. (See also [Hanks and McDermott 1985].) Further, for most actions the set of qualifications or possible thwarting circumstances is very large, if not infinite, so unless one rules out these exceptional cases, one cannot conclude anything at all useful about the effects of the action—one gets at best an intractably large disjunction. McCarthy [1977] has christened this difficulty the *qualification problem*, and proposed that it be solved by *circumscribing* the current beliefs about the action; this means making the general assumption that nothing untoward happens except explicitly expected exceptions, thus achieving a principled sort of ignorance about exceptions. (See also [McCarthy 1980].)

But there is another serious problem. McCarthy and Hayes ([1969] and [Hayes 1973]) observed that if the effects of actions are fully axiomatized, one needs to state not only those things that change, but also endless details about what does not change. They labeled this difficulty the *frame problem*. What seemed needed was some way of stating the few obvious changes explicitly and making as few unstated changes as possible in accommodating the stated changes (see [Shoham 1987]). McCarthy [1986] attempts to address the frame problem by using his notion of logical circumscription to say that the only changes are those explicitly derivable from the descriptions of axioms, and that if a possible change is underivable it is assumed not to occur. Computational systems like STRIPS finessed this problem by using a mutable database to store their beliefs. Such systems simply made a few additions to and subtractions from the database to account for the explicit action effects, and carried along all other beliefs unchanged. Waldinger [1977] called this the "STRIPS assumption," but it is merely a special case of the older philosophical idea of conservatism.

2.4.1 Conservative accommodation

Most theoretical prescriptions in philosophy and actual practice in artificial intelligence restrict the admissible changes of state to *conservative* ones, changes which keep as much of the previous state as possible. (Quine [1970] uses the term "minimum mutilation" for the same notion. See especially [Gärdenfors 1988]; also [Quine 1953], [Quine and Ullian 1978], [Ellis 1979], and [Harper 1976].) For example, in addition to the STRIPS assumption, each of the backtracking procedures used in artificial intelligence represents some notion of minimal revisions. In "chronological" backtracking, the agent keeps all beliefs

except the ones most recently added. "Non-chronological" or "dependency-directed" backtracking is even more conservative, abandoning as small a set of beliefs as possible regardless of the order in which they were adopted. For instance, the procedure for dependency-directed backtracking given in [Doyle 1979] minimizes the changes by abandoning only "maximal assumptions." (See also [Reinfrank 1985], [de Kleer 1986], and [Goodwin 1987].)

The general notion can be made precise in the following way. (See also [Gärdenfors 1988] and [Fagin et al. 1983].) Conservatism supposes the existence at each instant of a comparison relation between states, a way of comparing the relative sizes of changes entailed by moves to different states. For $X, Y \in \mathcal{I}$ we write this relation $X \underset{t}{\preceq} Y$, read as "$X$ is at least as close to S_t as is Y," and require that $\underset{t}{\preceq}$ be reflexive and transitive; that is, that $\underset{t}{\preceq}$ be a quasi-order. With these instantaneous comparison relations, we define $\gamma_t \subseteq \Gamma_t$, the set of minimal changes or nearest successors at instant t, to be those successors minimal under $\underset{t}{\preceq}$, that is,

$$\gamma_t = \{S \in \Gamma_t \mid \forall S' \in \Gamma_t \quad S' \underset{t}{\preceq} S \supset S \underset{t}{\preceq} S'\},$$

and say that the change from S_t to S_{t+1} is conservative if $S_{t+1} \in \gamma_t$. (A formally similar notion appears in logical treatments of counterfactuals, where the quasi-order is called a comparative similarity relation. See [Lewis 1973].) There need not exist any conservative change if Γ_t has no minima under $\underset{t}{\preceq}$ (if there is an infinite descending sequence, for example).

Formulated this way, conservatism involves measures of the closeness of states and minimality of changes. The definitions do not restrict the choices of such measures in any substantive way—for example, the weakest comparison relation $\underset{t}{\preceq} = \mathcal{I} \times \mathcal{I}$ is admissible, and under this relation every change is minimal, hence conservative—so each concrete application of this notion must supply its own specific definition. For instance, the conservatism of the dependency-directed backtracking system DDB presented in [Doyle 1979], which is a complicated sort of conservatism distinguishing "premises" from auxiliary "assumptions," embodies a measure of changes in terms of the state differences they represent. In this case, the comparison relation compares sets of differences from the initial state, that is, the sets of added and deleted elements:

$$A \underset{t}{\preceq} A' \text{ iff } A_t \triangle A \subseteq A_t \triangle A'.$$

($A_t \triangle A$ is the symmetric difference of the sets A_t and A, that is $(A_t \setminus A) \cup (A \setminus A_t)$.)

A related measure, which Harman [1986] calls the Simple Measure, compares instead the cardinality of these sets:

$$A \preceq_t A' \text{ iff } |A_t \triangle A| \leq |A_t \triangle A'|.$$

2.4.2 Rational accommodation

Even though the conservatism of DDB exhibits an appealing formal simplicity, it is unsatisfactory for practical use except in simple cases that do not depend much on conservatism for success. The same holds as well for the other "blind" backtracking systems widely used in artificial intelligence. Both of the specific comparison relations defined above treat all beliefs equally, and are as ready to discard eternal truths as they are to discard rumor and raving. They are not unreasonable comparison relations for systems like DDB that manipulate representations of beliefs without regard to the meanings of those beliefs, but the cases of interest in scientific and mundane reasoning (whether human or machine) call for conservatism to respect the differing values of beliefs of differing content, to not consider all consistent sets of beliefs equally acceptable. In meaningful revisions, the comparison measure compares the values of the beliefs adopted and the beliefs abandoned; one state will rank closer than another if it abandons less valuable beliefs or adopts more valuable ones. These values represent preferences about possible revisions.

But if conservatism is to respect preferences about possible changes, whose preferences are respected? In the case of DDB, the system, through its very construction, respects the preference of its designer for minimal sets of changes, a preference which is not explicitly represented anywhere in the system's program. One might redesign DDB to include a big table stating the value of every possible statement, and to consult the table when revising beliefs—roughly the approach taken in [Rescher 1964] and in [McAllester 1980]. But it is more natural to think of belief revision as a sort of action the agent takes, guided, as with its other actions, by its preferences, in this case preferences about which beliefs it should hold, preferences about its own nature. We call this *rational* accommodation. For example, in our own (human) reasoning, the preferences guiding belief revision are often our own. Our greater credence in Newton's laws than in our neighbor's gossip is not something inherent in our construction (though there may be innate features of our reasoning). That greater credence reflects the values we have developed or learned.

Conservative accommodation is one of the principal forms of rational accommodation. Rational volition aims at improving the agent's situation, at maximizing its satisfactions while minimizing its dissatisfactions. We might call this reasoning from states to intended changes *progressive* reasoning, the intended improvements representing progress for the agent. Complementing this, rational accommodation aims at effecting these improvements efficiently. Since the unintended changes were not rationally chosen in rational action, they represent non-rational or irrational changes, so rational accommodation seeks to maximize the rationality of the agent by minimizing the irrational component of changes. The main sort of irrationality to be avoided is abandoning the wrong attitudes, so we may think of conservatism defined in terms of the agent's values as *rational* conservatism.

In rational conservatism, states are compared by means of the preferences about states they contain. If states express weak preferences over all possible states, conservatism may maximize the rank (utility) of revisions by defining $X \underset{t}{\prec} Y$ iff $\{Y\} \lesssim \{X\}$ in the preference ranking according to the states involved at instant t. Here we have a choice of two forms of conservatism, which we call *predicted* conservatism and *ratified* conservatism. In predicted conservatism, the preferences $\pi(S_t, S_t)$ of the current state S_t are used in comparing the possible successors, so that at each instant the agent predicts which successor will be most preferred. In ratified conservatism, the preferences of each possible successor are used to compare the alternatives, and only successors that rank themselves at least as highly as they rank all other successors represent conservative choices. That is, $S \in \Gamma_t$ is a conservative successor only if $S \in \gamma_t$ when the relation $\underset{t}{\prec}$ used in defining γ_t is determined by the preferences in S, that is, by $\pi(S, S)$. In this case, each conservative successor ratifies its own choice. (The term "ratifies" is borrowed from Jeffrey's [1983] similar notion of ratified decision.)

2.5 Reflection

The preceding has introduced the basic concepts of rational self-government, but as relatively abstract notions. In this section we turn to applications, using the concepts of rational self-government to reformulate, understand, and criticize some ideas and techniques in artificial intelligence.

The first observation one makes upon surveying the field is that most systems developed in artificial intelligence are pretty irrational, as they lack any

way of sensibly allocating the effort they expend in reasoning. Rational allocation of effort means not wasting effort making easy but worthless inferences, or on easy but unlikely steps, but instead focusing on steps of high expected value. Schemes for simply cutting off effort after some arbitrary number of steps of calculation are irrational, for they make success in reasoning unrelated to subject or knowledge. Yet artificial intelligence has relied heavily on arbitrarily imposed limits to calculation. Similarly, domain-independent methods of automated deduction are irrational, since worthwhile inferences in one domain may be of the same logical form as worthless inferences in another. Despite all the effort expended to make deduction of specific conclusions as efficient as possible, in these logical systems, interest, desire, and likelihood are missing: logically, every conclusion is equally interesting and every path equally likely. Because these systems operate in the same way no matter what the subject, they cannot be very discriminating or very efficient in their activities.

It is effort wasted in these ways that leads to senseless lapses of reasoning ability. Even when adequately knowledgeable about a subject, limiting reasoning in arbitrary and senseless ways yields ragged and unstable reasoning powers that make success almost accidental and performance unpredictable. The hallmark of current systems is that they can miss obvious inferences in a haphazard fashion, succeeding on one problem but failing or making errors on simpler ones or seemingly identical ones that few people would distinguish. Of course, whether one statement follows from others is in general undecidable, but even when attention is restricted to very limited, decidable classes of implications, the behavior of current systems is not accurately predictable. Even in simple cases, the only way to tell if a system will succeed is to try it and see, and much of the heavy emphasis on implementation in artificial intelligence stems from this unfortunate state of affairs. The ill-considered reasoning so common in current systems means complete but unreliable behavior, complete in the sense that some action is always taken. What is needed instead is well-considered reasoning and action, which is incomplete (sometimes nothing is done) but reliable (when something is done). As in human conduct, sensibly conducted reasoning may fail and make mistakes—that is unavoidable without omniscience and omnipotence—but none of the errors will be stupid, and that is the best we can expect, and is better than current practice in artificial intelligence.

To expend its effort wisely, the reasoner must possess knowledge of the limits of its own knowledge and abilities, and this means these limits must be regular or reliable enough to be generally describable or predictable by the

agent as well as by observers. But more than that, the reasoner must weigh the relative values of different possible inferences about a subject, so as to guide its reasoning in sensible ways. This evaluative knowledge, which may vary from subject to subject, is every bit as crucial to success and reliable performance as declarative knowledge. Its systematic neglect in the codification of subjects, visible in the blinding preoccupation with logical axiomatizations and deductive inferences, has spelled failure for many approaches to knowledge representation (see [Minsky 1975], [de Kleer, et al., 1977], [McDermott 1987], and [Winograd and Flores 1986]). As a consequence, most expert systems constructed to date are not artificial intelligences, but instead *idiot savants*, knowledgeable about specific subjects but knowing nothing else. These systems have proven very valuable in some applications, but in delicate situations this narrowness is cause for concern. Because these artificial idiots possess little or no understanding of the limits of either their own reasoning powers, knowledge about their subject, or knowledge about the context of their actions, they think and act blindly, using automatic procedures which mechanically apply whatever subject knowledge the system possesses to arrive at a conclusion. In contrast, human experts know that their function is not simply to solve problems or to make decisions, but to prudently *manage* the use of their knowledge and skills to solve problems and make decisions, taking into account their own powers, limitations, and reliability. Human experts not only know their subject, they know how to think about it—that is, they know when they can solve a problem and when not (generally without actually trying it to see), and whether these judgments can be trusted (see [Schön 1983]). In circumstances where human experts grow uncomfortable and suspect problems beyond their ken, the automated systems, as a result of their narrowness, simply act as usual. They never step back and say "I don't know what's going on—get someone who does!" In consequence, great care must be exercised by designers to ensure that the environment in which an expert system is used can tolerate its unwitting failures.

Let us call these two sorts of knowledge *direct* subject knowledge and *ratiocinative* subject knowledge, recognizing that different subjects may call for different ways of thinking—different heuristics, for example. (Following Polya [1962, p. vi], we call some sorts of ratiocinative subject knowledge—"the means and methods of problem-solving"—heuristic. The distinction between direct and ratiocinative subject knowledge goes by other names as well: declarative and procedural, epistemological and heuristic (in artificial intelligence), theoretical and practical knowledge (in philosophy), and knowing what and knowing how

in ordinary discourse.) Indeed, the bodies of direct and ratiocinative knowledge about one subject are subjects in their own rights, and if sufficiently complex may call for further direct and ratiocinative knowledge about these new subjects.

While ignorance and error are unavoidable, properly applied to a problem, direct and ratiocinative subject knowledge can yield not mere actions and conclusions, but, as with a human expert, reflections on their accuracy, completeness, appropriateness, and reliability—in sum, *considered* actions and conclusions. Through self-evaluative reasoning, we may still seek what the Greeks called *sophrosyne*. In today's English, sophrosyne is usually used as the antonym of hubris, meaning temperance or moderation. But sophrosyne (to quote from [Ostwald 1962, pp. 313-314]), "literally translated, means 'soundness of mind,' and describes the full knowledge of one's limitations in a positive as well as a negative sense: the *sophron*, who possesses this virtue," "knows what his abilities and nature do and do not permit him to do. He is a self-controlled man in the sense that he will never want to do what he knows he cannot or should not." "Though self-control is more negative than positive in modern usage, if the word is taken more literally than it usually is, i.e., if 'control' is not merely taken as 'restraint' but also as 'mastery,' it comes closer to sophrosyne than most alternative renderings."

There have been several steps taken towards rational self-government in artificial intelligence. Most of these have expressed the idea in different ways, including, for example, control of reasoning, meta-reasoning, reasoning about reasoning, meta-planning, reflection, and introspection. One of the earliest proposals along these lines was by Hayes [1974]; other discussions include [McDermott 1978], [de Kleer, et al. 1977], [Davis 1980], [Weyhrauch 1980], [Doyle 1980], [Smith 1982], [Stefik 1980], [Smith 1985], [Lenat et al. 1983], and [Laird et al. 1987].

In the rest of this section we survey some of the issues and approaches in the context of some fundamental applications of rational self-government. Each application involves reflection on one or more sorts of the agent's attitudes. Assumptions and learning involve reflecting on one's beliefs; planning involves reflecting on intentions and priorities; and deliberation involves reflecting on one's preferences and desires. Unfortunately, discussions of these activities in artificial intelligence usually downplay the strong sense of unity they exhibit. In part, this has occurred because the field of artificial intelligence has focused on the notion of knowledge (whether direct, ratiocinative, or self knowledge) to the virtual exclusion of the notion of rationality. Since there are many domains

of knowledge (both expert and common), many completely separate from each other, this focus has produced a very disjointed field, with many researchers having little in common besides implementation tools like LISP or PROLOG. Even commonality in language of expression is not assured, since different schools focus further on restricted forms of knowledge: logicists focusing on factual knowledge, search theorists focusing on evaluative knowledge, and proceduralists focusing on procedural knowledge. The only way of unifying these disparate sorts of knowledge is through the forgotten ideal of rationality, which provides a common intellectual framework with which to frame, communicate, and use knowledge in thinking. While our knowledge of separate subjects may be pursued separately, the notion of rationality is central in unifying the use of knowledge in the activities of reasoning, searching, planning, problem-solving, and learning. It is therefore crucial in making artificial intelligence intelligible to readers from other fields, such as philosophy, logic, statistics, decision theory, and economics. When artificial intelligence removes rationality from thinking the fragmentary topics and techniques that remain seem unmotivated and unrelated to outsiders for whom rationality is a central concept. (See also [Doyle 1988b] and [Miller 1986].)

2.5.1 Assumptions

Thinking often begins with making guesses grounded in one's experience. Guessing, or making assumptions, is often held in disrepute as illogical. In fact, though illogical, it is often quite the rational thing to do. Taking action requires information about the available actions, about their expected consequences, and about the utility of these consequences to the agent. Ordinarily, obtaining such information requires effort, it being costly to acquire the raw data and costly to analyze the data for the information desired. To minimize or avoid these costs, artificial intelligence makes heavy use of heuristics—rules of thumb, defaults, approximately correct generalizations—to guess at the required information. These guesses are cheap, thus saving or deferring the acquisition and analysis costs. But because they are guesses, they may be wrong, and so these savings must be weighed against the expected costs of making errors. Most of the cases of default reasoning appearing in artificial intelligence represent judgments that, in each particular case, it is easier to make an informed guess and often be right than to remain agnostic and work to gather the information; that errors will be easily correctable and ultimately inconsequential; and that the true information

needed to correct or verify these guesses may well become available later anyway in the ordinary course of things. In other cases, defaults are avoided, either because there is no information available to inform the guess, or because even temporary errors of judgment are considered dangerous. These ratio-economic judgments may also be influenced by non-economic desires and preferences, such as moral or ethical judgments of positions ("assume people innocent until proven guilty"), and social conventions for cooperation and communication (assume dumb questions aren't, so that, for example "Can you pass the salt" means "Please pass the salt").

Rational inculcations of beliefs have been recognized for many years, famously by Pascal [1662] and James [1897] in the context of religious belief. Pascal, for example, framed his problem of belief in God as the following: he can either believe or doubt the existence of God, and God may either exist or not exist. If God exists and Pascal believes, he gains eternal salvation, but if he doubts he suffers eternal damnation. If God does not exist, belief may lead Pascal to forgo a few possible pleasures during his life that doubt would permit him to enjoy. We may summarize these evaluations in a decision matrix

Pascal's decision	God exists	doesn't
Believe	$+\infty$	$-f$
Doubt	$-\infty$	$+f$

where f represents the finite pleasures enjoyed or forgone during Pascal's life. Of course, these same quantities modify the first column as well, but finite modifications to infinities are negligible. As long as God's existence is not judged impossible, the expected utility of belief is $+\infty$, dominating the expected utility of doubt, $-\infty$.

But the same sorts of rational revisions are ubiquitous in mundane reasoning, and many of the common sorts of apparently non-logical reasoning studied in artificial intelligence reflect the results of such economic calculations. For example, in the morning my habit is to get in my car with my notebooks and start the car, in order to drive into work. Now the car might either be working or broken. It must be working for me to be able to use it to drive to work, but I do not check to see that it is before trying to start it. I simply assume it is working when I plan and pack the car. We can frame my decision in a matrix

My decision	car works	doesn't
believe	$+b - c$	$-c - C$
doubt	$+b - C$	$-C$.

Here we write the benefit from the car starting as b, the cost in effort of packing and starting the car as c, and the cost of checking out the engine, electrical system, transmission, etc. as C, where we assume $C \gg c$. With these utilities, the expected value of believing dominates that of doubting whenever $pC > c$, where p is the probability that the car works. As long as I expect the car to work and $C \gg c$, my assumption is reasonable.

This sort of economic calculation may be made either at the time the information is needed or, as in the default rules prominent in inheritance systems and reason maintenance, in advance. (See [Smith 1985] and [Langlotz, et al. 1986] for detailed treatments of these ideas.) When made in advance, the rules of thumb may be applied to get the guesses either when needed, as in most inheritance systems, or in advance, as in reason maintenance. In most current artificial intelligence systems, however, these calculations are made by the system's designer—the human informants decide what the good guesses are, and these are encoded into the rules that the machine obeys. But these considerations may be made by the agent itself as well through reflection and reasoning. To judge the benefits of making some guess, the agent might assess the impact of the guess on its state of mind—that is, ask whether the assumption really matters, or whether it would not change the agent's effective knowledge very much. It might assess the expected savings—will it have to do the work anyway?—and assess the costs of taking a position—will making any guess get it into arguments or explorations that will take more time than it is worth? (See [Elster 1979, 1983], [Brunsson 1985], [Pears 1984], [Hirschman 1982], and [Levi 1967, 1980].)

There have been attempts in the artificial intelligence literature to view heuristics or rules for making assumptions purely in probabilistic terms, with a rule of assumption justified as long as its probability exceeds some threshold value. This theory of assumptions is inadequate. Tautologies have maximum probability but are generally worthless. No agent should waste its effort assuming most tautologies, since tautologically, an assumption is worth the expense of making it only if it is worth it—that is, if the expected utility of making it is high enough. Since the probabilistic theory of assumptions ignores the utility or disutility of assumptions, it is a theory of likely irrelevancies, of tasteless

theorizing.

A similar, but less popular mistake is to base assumptions purely on utilities, assuming something as long as its utility exceeds some threshold, regardless of the probability of its being true. This theory has exactly the same irrational character as the probabilistic theory of assumptions, and has a standard name as well. It is called wishful thinking.

2.5.2 Learning

Rationality enters learning in the rational selection of what to learn or investigate, what to assume, what to consider, and what sorts of conclusions to seek. The first choice is that of subject, of the aim of learning. Because they are silent on this choice, most treatments of learning in the philosophy of science, inductive logic, and artificial intelligence seem terribly bloodless, if not misguided. They study learning that has no point, learning irrationally pursued towards no end. In everyday life, things do not just suggest themselves as worth learning. In most cases outside of motor skills and perhaps infant language learning, one learns (towards specific or abstract ends) because one wants to, and regularly fits new information into theories of current interest. When the will to learn ceases, the learning stops, even if the agent later suffers for his blindness. But in most work in artificial intelligence, learning involves no errors, no selection, no choice of vocabulary—no ends. The designers of the systems supply algorithms, vocabularies, and data, and the program mindlessly applies the algorithm. These systems "learn" things in exactly the same sense that a bank's computers "learn" the new balance in an account after some transaction. True learning essentially involves choice, and no learning occurs without ends, only computation. Perhaps work on aimless learning in artificial intelligence should not be faulted much, for as Truesdell [1984] points out, most theories in the philosophy of science are similarly aimless, and the ideas on learning in artificial intelligence have been strongly shaped (sometimes unwittingly) by theories in inductive logic and the philosophy of science. (See also [Rorty 1979] and [Grabiner 1986]).

Once a subject is chosen for learning, the agent must choose how to gain information about or experience with the subject. In some cases, the aim will be to learn something from what the agent already knows or has experienced, so that no additional investigations need be conducted. In cases in which the agent or its designer decides (or is forced to conclude) that real effort must be made to acquire and analyze information instead of just guessing, economic calculations

may guide the processes of acquisition and analysis. Data may be gathered in many ways—by exploring the world, by exploring the agent's current behavior, or by enrolling in classes, reading books, and asking questions—and the agent must choose among these methods.

But once these decisions are made (and they are subject to repeated revision, of course), the biggest problem in learning is that alternative explanations must be compared. The grounds of these comparisons are naturally viewed as reflecting the preferences of the agent about what explanations are better than others. These comparisons occur not just in the initial presentation of alternative theories, but in the course of the investigation as well, when one must adapt one's explanations, hypotheses, and theories to new information. This is especially visible in the approaches to learning via analogies and metaphors between cases. In making analogies and interpreting metaphors, one must deform one explanation or concept into another, somehow judging one deformation better, milder, or more apt than another. This task is substantially the same as that of accommodating changes of attitudes, where the different aspects of the new case indicate the stipulated changes. Similarly, in improving one's skills, one must choose among alternative explanations of the flaws in one's skills (a choice called "credit assignment" in artificial intelligence), and also choose among alternative patches or modifications to one's skills. As in accommodation, preferences about revisions are one natural way of formulating these judgments. (See [Carbonell 1986].)

One important special case of such comparisons is comparison of alternative formulations or conceptual organizations on economic grounds. Every formulation entails costs in memory and time: the memory needed to store the information, and the time needed to use it. There are well known tradeoffs between the succinctness of axiomatizations and lengths of proofs, and between expressiveness of languages and difficulty of finding proofs. These same economic tradeoffs seem to motivate some sorts of learning, such as seeking simple descriptions of lots of data in order to compress it. More importantly, such economic judgments seem to motivate organizations of concepts into hierarchies of prototypes, as is common in both human thought and artificial intelligence. Hierarchies, of course, may offer dramatic (often exponential) economies of storage space. Prototypes, or natural kinds, offer related economies. For example, say a prototypical concept is defined as the conjunction of n properties or aspects, and that objects satisfying any $n - 1$ of these properties are counted as instances of the concept, albeit exceptional ones. Suppose further that we wish to describe

$n+1$ individuals, one of which satisfies the concept perfectly, and n exceptional instances representing every possible exception. To describe these individuals in a system of prototypes and exceptions requires only $2n+1$ statements: a typing or IS-A statement for each instance ($n+1$ all together) and a statement of the exceptional property for each of the exceptional instances (n of these). But to make the same descriptions using the ordinary logical connectives and implication requires n^2+1 statements: one implication for the perfect instance, and n statements, variations of the prototype's definition, to describe each of the exceptional cases, n^2 in all. (See [Lakoff 1987] for more on conceptual hierarchies.)

Similar economic judgments, this time about the cost of retrieval instead of the cost of storage, motivate the use in artificial intelligence of redundancy. This means using several distinct paraphrases or reformulations of the subject knowledge in hope that these "multiple perspectives" increase the chances of solving the problem by permitting more possible successful derivations, some shorter or easier to construct than is possible with any single organization of the knowledge (see [Sussman and Steele 1980]). But additional paraphrases also make for more possible unsuccessful derivations, perhaps increasing rather than decreasing their relative frequency, and one must judge in each case whether the expected benefit is positive or negative.

Economic comparisons of theories in terms of the time needed to use them are at the heart of what is called heuristic. Artificial intelligence makes much of the notion of heuristic without saying precisely what this notion is. One view, natural in the present setting, is that heuristics are rules or algorithms for reasoning that trade off accuracy and certainty for speed and cheapness of use. Unfortunately, examples of analysis of heuristics in artificial intelligence are rare, but see [Langlotz, et al. 1986], [Knuth and Moore 1975], and [Pearl 1984].

One sort of heuristic is that captured in the notion of probabilistic algorithm. When it is too hard (undecidable or intractable) to reason exactly towards a solution, one response is to guess the answer. To be useful, the expected correctness of the guess should be high, and the expected cost low, so that the expected value of the guess is high. Probabilistic algorithms are algorithms that work by guessing, that is, algorithms that compute answers which may be right or wrong, but which are right over half the time. Repeated applications of such algorithms can re-check the answer, thus achieving any specified degree of certainty. Very good probabilistic algorithms are known for message routing, primality testing,

and other problems for which the best known exact solutions are quite bad. See [Harel 1987], [Cook 1983], and [Karp 1986] for more on these. In the context of learning, probabilistic algorithms occur most naturally in the form of probabilistically induced concepts. See [Valiant 1984] for a discussion of this approach to learning.

Another sort of heuristic is that captured in approximate algorithms, which compute answers that may not be entirely correct, but are good enough. Simon [1969] calls this *satisficing*. While in probabilistic algorithms one chooses the desired frequency of correct answers and then seeks an algorithm yielding these, in approximation algorithms one chooses the acceptable degree of error in the answer, and seeks an algorithm yielding this. The traveling salesman problem, for instance, has no known fast exact solution, but admits fast approximation algorithms whose answers are within a factor of two of the correct answer. In fact, the concepts induced by Valiant's [1984] learning algorithm also qualify as approximate algorithms in this sense.

The use of probabilistic and approximation algorithms involves a rational decision to trade off certainty and accuracy for speed. But in these cases, the bargain one makes is well understood, in that the risks of error and inaccuracy can be quantified, indeed, chosen by design. Artificial intelligence makes heavy use of less well understood heuristics, of rules for making cheap guesses which are only hoped, not known, to help. For instance, one employs methods which one subjectively expects to be likely correct or only slightly inaccurate, even when these bounds themselves are mere guesses, not guarantees. In really difficult problems, trying all the heuristics one can think of is all one can do, for to wait for an exact solution is tantamount to giving up. The danger, of course, is that these informal heuristics might actually hurt more than they help. But as long as they are cheap to make, and their errors easy to correct, one might as well use them. Especially in ill-understood problems, applying heuristics often leads to the discovery of information which is correct and useful in other parts of the investigation, even though the heuristic fails to achieve its nominal purpose.

2.5.3 Planning

Where assuming and learning involve selection, formulation, and revision of beliefs, planning consists of selection, formulation, and revision of the beliefs, intentions, and priorities relevant to one's activities. These include beliefs about the effects of actions (so that actions may be selected as methods for achieving

specific effects), beliefs about what circumstances will obtain in future situations, intentions about what actions will be performed, intentions about how these will be performed, and priorities influencing the temporal order in which these intentions will be carried out. The steps of forming and revising these intentions and priorities resembles the steps of forming and revising beliefs in learning in that the agent reflects on the consistency and completeness of these attitudes.

The formation and revision of absolute intentions involves deciding what to do and what not to do. One source of such decisions might be to adopt intentions to satisfy the currently maximal desires. Another source is to consider whether one's plan to achieve some aim is complete. Here levels of abstraction may be considered one at a time to see if the planned actions at each level achieve their aim, or if they cannot be expected to. If not, new steps must be added. This operation, usually called goal regression, is one of the main functions of automated planners like NOAH [Sacerdoti 1977] and Waldinger's [1977] planner. Another source of intentions is judging the completeness of plans at different levels, that is, seeing if ways are known for carrying out intentions at one level with definite methods (see [Sacerdoti 1974] and [McDermott 1978]). This sort of completion of the plan need not be postponed until an intention has been selected for immediate action, but may be decided on in advance. Methods themselves may be either primitive actions from among the agent's current skills, or subplans that perhaps require further reduction themselves. One may also reflect on the consistency of one's intentions, or on the consistency of one's intentions with one's desires. If, for example, information newly learned shows two current intentions to be inconsistent, the set of current intentions might be revised to restore consistency, either by modifying or abandoning one of the conflicting intentions, or by adopting a priority so that they are not both considered for action at the same time. Similarly, if one strongly dislikes or comes to dislike the aim of some intention, one might simply abandon the intention.

Paralleling the formation and revision of absolute intentions, planning involves forming and revising priorities. Priorities for actions may be considered only at the last moment when unavoidably required in order to select the next intention to act upon in volition, or in advance. These priorities may be judged according to completeness and consistency as well. Deciding what to do next in volition is the most obvious case of making the set of priorities more complete, adding in first priority for the selected intention if the current set does not contain a first priority. Priorities also may be inferred from the intentions themselves, as when data or preconditions required by actions yield temporal

orderings among steps. Determination of such dataflow priorities is a central operation of automated planners like NOAH. Similarly, revision of the priorities may be necessary if new information indicates the current priorities to be inconsistent. This criticism and revision of orderings is also central to NOAH and HACKER [Sussman 1975]. A more subtle version of consistency checking is that of ensuring intended conditions which must be preserved throughout a sequence of steps. Such path conditions may restrict the possible orderings. If they restrict them too much, an inconsistency results and the conditions must be revised. See [Chapman 1987] for a general planning algorithm addressing many of these issues.

2.5.4 Deliberation

When reasoning concerns the agent's preferences as well as the agent's beliefs, we call the process deliberation. In ordinary decision theory, the agent's preferences are taken as givens of the decision-making process. But in ordinary life and in artificial intelligence, this role for preferences is too restricted, for in many cases one must perform some amount of reasoning to determine which of one's preferences apply to a particular decision, or more fundamentally, what one's preferences are regarding some heretofore unconsidered possibility. Filling in incompletenesses in one's preferences and resolving inconsistencies between one's preferences mean that deliberation involves selecting or choosing (and possibly inventing) a set of preferences relevant to a decision problem. Together with reasoning about what alternatives exist and what their consequences are, this reasoning about what preferences to use constitutes the process of deliberation.

An abstract procedure for deliberation might go as follows (see [Doyle 1980]). In this procedure, the agent constructs a set of attitudes called the *decision situation.* The decision situation contains the set of attitudes decided to be relevant to a particular decision problem, and so embodies the agent's current formulation of the nature of the decision. Initially, the decision situation may be empty, but then the agent iteratively makes incremental revisions of the decision situation. The agent repeatedly searches its attitudes for beliefs about what are possible options or alternatives for the decision problem; for likelihoods about their consequences; and for preferences among these. Each time some attitude is added to the decision situation it is scrutinized to see if it is truly applicable to the problem. This involves checking for exceptional features of the problem and formulation that defeat the likelihood or preference in this case. Ordinarily,

the amount of information possessed by the agent about explicit exceptions to a particular attitude is very limited, so this routine criticism need not be carefully controlled. But the set of attitudes possibly relevant to the decision problem is usually very large, and most of the rational control of the progress of deliberation must concern whether to seek further relevant information, and whether to add particular attitudes to the decision situation. One element of this control or second-order decision of whether to pursue deliberation is to judge the properties and import of the current formulation. If the likelihoods and preferences currently identified are inconsistent, some resolution must be decided on, using information about exceptions or using preferences about preferences and preferences about likelihoods to defeat or discard some of the attitudes and produce a coherent view of the decision problem. If the likelihoods and preferences are too incomplete to narrow the choice sufficiently, more may have to be sought. Once a consistent formulation is achieved, the agent may see what outcomes it entails, and with what degrees of confidence. The agent may then decide whether to stop or continue based on its feelings about risk, about the stability of the decision (whether unconsidered information is likely to change the outcomes significantly), about the number of acceptable options, and about the outcomes themselves. For example, if it very much desires one tentatively selected alternative, it may do unusually little work to investigate other alternatives and stop prematurely; if it very much dislikes some outcome it may seek new alternatives or new arguments to prefer other alternatives more. Similarly, if the agent desires certainty or prefers less risk, it may prolong reasoning in order to increase certainty about some iffy conclusion, or curtail reasoning that involves questioning of favorite dogmas.

2.5.5 Introspection

Learning, planning, and deliberation involve reflection on and revision of most of the attitudes of action: beliefs, likelihoods, preferences, intentions, and priorities. But they do not necessarily involve forming or revising desires, or the more fundamental sorts of beliefs and preferences. Indeed, most of life changes desires only indirectly, if at all, for few people really spend much time reflecting on what they think they should want. When disaster or disappointment do lead us to wax philosophical and consider the meaning of life and how we should live, we are usually timid in the depth of self-criticism we are willing to consider—rightfully so, for fundamental change is dangerous. Once one starts changing

one's fundamentals, anything can happen. Nevertheless, self-examination is often very useful. Socrates' belief that the unexamined life is not worth living still holds true, despite the usual victory of Gresham's Law.

Disappointment is perhaps the most common cause for introspection into and revision of desires. The limits on our knowledge of the world extend to limits on our knowledge of ourselves, and it is commonplace that we do not understand many of our desires and preferences very well, even though we bemoan other people not understanding us either. Worse, in some cases we mistakenly think we understand them, and use these mistaken views in deliberating. How familiar it is for the child to strive a lifetime to satisfy the desires of the parent, only to discover in the end how foreign and unsatisfying these accomplishments are. (See [Scitovsky 1976], [Hirschman 1982], and [Schelling 1984b] for more discussion.)

Living without self-omniscience is characteristic of human life, and conceivably this circumstance might be avoided in the design of artificial agents. In humans it appears intimately tied up with the modular nature of mind, the "social" brain or mind as Gazzaniga [1985] and [Minsky 1986] put it. It may be that thinking demands psychological organizations of sufficient complexity to ensure that similar troubles afflict any intelligence (see [Thomason 1979]), but that is a question for investigation.

Even with perfect self-omniscience, fruitful opportunities remain for reflection on desires. Most of these concern the economic and moral consequences of one's desires. One might desire quiet living, but also recognize the extreme cost it entails in terms of one's other desires. One might then live by frustrating one desire or another. Alternatively, one might increase one's happiness by abandoning one of the conflicting desires, so as to be able to satisfy the remainder without frustration. People would not be human without many conflicting desires, but it is always a choice whether to suffer or to remedy particular conflicts (see [Peck 1978]). The Buddha said to abandon all desires to avoid suffering, and that is a sufficient means. But piecemeal revision is the ordinary path to enlightenment and equanimity, and total selflessness may not be necessary means to happiness. Rawls [1971], for example, introduces the idea of *reflective equilibrium*, states of mind in which all of one's attitudes and principles are in harmony. The Buddha's nirvana is one such equilibrium, but there may be others as well.

Chapter 3

Constitutional self-government

Not all reasoning need be made deliberately through volitional procedures. Some reasoning may be automatic, made without specific motivation in supplying a sort of "background," "unconscious," or "common sense" reasoning for deriving "obvious" consequences of one's attitudes. This sort of reasoning figures in Harman's [1986] notions of immediate implication and immediate inconsistency,[1] and occurs prominently in most artificial intelligence systems in the form of restricted sorts of inferences the system always makes by using special-purpose procedures or machines, for example, the inferences made by inheritance networks ([Fahlman 1979], [Touretzky 1986]), by reason maintenance systems ([Doyle 1979], [de Kleer 1986]), or logically obvious inferences ([McCarthy 1958], [Davis 1981], [Levesque 1984]). We call this sort of reasoning *constitutional* reasoning since it derives from the makeup or constitution of the agent, independent of its specific plans.

The most important form of constitutional reasoning is the unintentional, accommodative reasoning expressed in the state space $\mathcal{I}$. When one chooses a state space for an agent, one also chooses a measure of automatic reasoning, since making only intended changes in a state may yield a set of attitudes outside the chosen state space. If $\mathcal{I} \neq \mathbf{P}\mathcal{D}$, the modified set of attitudes may not be in $\mathcal{I}$, so to accommodate the intended changes some additional, unintended changes must be made to yield a state in $\mathcal{I}$. Especially when there is a single possible accommodation, and hence no choice involved in accommodating the intended effects, we think of these additional changes to states as automatic reasoning.

State spaces are ordinarily defined or constructed by imposing conditions on

[1] See Appendix A for more on these.

the constitution of the agent, for example, restrictions to consistent or suitably complete sets of attitudes. We call such conditions and restrictions *constitutive assumptions*, assumptions about or stipulations of the agent's structure. In the following, we will consider two sorts of constitutive assumptions: *laws of thought*, which are individual rules for self-regulation that the agent adopts, and *psycho-logics*, which are underlying logics of attitudes that specify the minimal consistency and closure properties of the agent's states.

3.1 Laws of thought and legal states

While it is possible to consider ideally rational agents with full logical powers, in which all logical entailments and consistency are automatic, such automatic reasoning is unmechanizable, so artificial intelligence pursues more limited sorts of automatic inference. The natural candidate for automation is common sense reasoning. In the usual, but rarely examined conception, common sense reasoning is the easy, obvious, uncontroversial, effortless, unconscious reasoning that humans share and that they use to support more difficult conscious reasoning.

Traditionally, the heart of commonsense reasoning has been logic, in its role as the laws of thought. Unfortunately, logic suffers several severe flaws as a candidate for automatic reasoning. As noted above, full logical deduction and consistency are unmechanizable. But as Harman [1986] points out, even when one restricts attention to "obvious" logical principles like Modus Ponens and Non-Contradiction, one faces the problem that repeated applications of obvious principles yield non-obvious results, so that requiring states of belief closed under Modus Ponens and satisfying Non-Contradiction may be as infeasible as requiring full logical omniscience. More telling, however, is the non-logical nature of common sense reasoning. As Harman also points out, implications like "if today is Tuesday, tomorrow is Wednesday" are as obvious and common in human reasoning as are tautologies—indeed, usually more obvious than most tautologies to most people. If non-logical inferences reflecting these implications are to be part of common sense reasoning, then common sense reasoning cannot be simply logical. Moreover, these non-logical rules of common sense reasoning vary from culture to culture, so that common sense is not a universal notion among humans (see [Geertz 1983]). Logic was initially conceived as the universal laws of thought, but a more reasonable view is that laws of thought, if they exist, are neither universal nor logical, but local instead. (See

also [Thomason 1987].)

3.1.1 Constitutive intentions

We here consider an extremely local conception of laws of thought, in which laws of thought are individual, with rules for self-regulation adopted by the agent as well as rules supplied by the agent's culture. Specifically, we conceive of laws of thought as *constitutive intentions*; or more specifically, as standing and routine constitutive intentions.

We classify intentions as standing or singular, routine or problematic, constitutive or environmental. Standing intentions are policies left in force and constantly obeyed until abandoned, while singular intentions are the ordinary sort abandoned once carried out. Routine intentions are intentions which the agent can satisfy directly through performance of one of its basic actions or skills, while problematic intentions require thought to carry out. Constitutive intentions are those strictly about the agent's own structure, while environmental intentions are those about the agent's environment or the agent's relation to it. Thus viewing laws of thought as standing and routine constitutive intentions (for brevity, simply constitutive intentions in the following) means that they are rules about the agent's mental structure that are always and immediately followed. For comparison, cases of rational assumption-making like Pascal's wager or James' will to believe reflect singular problematic constitutive intentions.

We identify the set of constitutive intentions as a subset $\mathbf{I}^* \subseteq \mathbf{I}^a$ of the set of possible intentions. Since we allow the meaning of an intention to vary with the agent's state and environment, the set of constitutive intentions may also vary. For simplicity, however, we will assume that constitutiveness of intentions does not depend on the specific state of the agent.

3.1.2 Satisfying states

The laws of thought observed by the agent thus determine a set of legal states of the agent, namely those states which satisfy the agent's self-regulations. Since each state may contain different constitutive intentions, the legal states are those that satisfy each of the laws they themselves contain, one state not being bound to observe laws it does not adopt. In this way, legal states exhibit something of Rawls' [1971] notion of reflective equilibrium, agreement between the agent's principles and attitudes.

As above, if $x \in \mathbf{I}^a$, the meaning of x is that the agent intends to act to make its world $W = (S, E)$ be one of the worlds in $\iota(x)$. If x is constitutive, the environmental portion of the proposition $\iota(x)$ is satisfied by any world, so the intention reduces to the condition that $S \in \mathbf{i}(\iota(x))$. We say that a set $X \subseteq \mathcal{D}$ is *satisfying* just in case it satisfies each of the constitutive intentions it contains, that is, if $X \in \mathbf{i}(\iota(x))$ for every $x \in X \cap \mathbf{I}^*$. Note that bigger sets may contain more constitutive intentions, and so be harder to satisfy. Our assumption, then, is that every internal state of the agent is satisfying.

We may divide the agent's constitutive intentions into two parts, one reflecting the agent's initial endowment of unadopted constitutive intentions, the other reflecting the agent's subsequent actions and legislation. We allow the possibility that legislation may amend the initial constitutive intentions.

3.1.3 Conditional attitudes

One of the most common sorts of constitutive intention employed in artificial intelligence is the *conditional attitude*. While unconditional attitudes include beliefs, desires, and intentions, in both absolute and relative forms, a conditional attitude $\kappa \Vdash x$, read "κ gives x," combines an attitude $x \in \mathcal{D}$ with an enabling condition κ. Such attitudes are interpreted as constitutive intentions of the agent to hold attitude x whenever condition κ obtains, and so are forms of the notion of indicative conditional discussed in philosophy and linguistics, that is, conditionals that signal an intent to revise one's attitudes upon learning the hypothesis, as in "If it's Tuesday, this must be Belgium" (belief), "The fruit salad will be fine, if it doesn't have grapefruit" (desire), and "Since it's raining, I'll take my umbrella" (intent). Conditional attitudes take many other special forms in artificial intelligence, including what are called "justifications," "reasons," "inheritance links," and MYCIN-style "production rules." See [Doyle 1982] for further discussion of the role of these intentions in artificial intelligence.

The enabling conditions of conditional attitudes are not themselves attitudes, but instead propositions about the state of the agent, in particular, propositions about that state of the agent that results from carrying out these constitutive intentions. Formally, we interpret κ as an internal proposition $\kappa \subseteq \mathbf{P}\mathcal{D}$, and define the internal meaning of $\kappa \Vdash x$ by

$$\mathbf{i}(\iota(\kappa \Vdash x)) = \{X \subseteq \mathcal{D} \mid X \in \kappa \supset x \in X\}.$$

Conditional attitudes can be automated only if the enabling conditions are

effectively computable. Artificial intelligence goes far with very simple sorts of computable conditions, namely those that refer to the presence or absence in states of specific sets of attitudes. The simplest sorts of these specify that the attitude x should be held if the state contains each of the attitudes in a specific set Υ^+ (the *in-hypotheses*) and none of the attitudes in another set Υ^- (the *out-hypotheses*). We write such an intention as

$$\Upsilon^+ \setminus\setminus \Upsilon^- \mid\!\!\vdash x,$$

read "Υ^+ without Υ^- gives x," and define its intentional content by

$$\iota(\Upsilon^+ \setminus\setminus \Upsilon^- \mid\!\!\vdash x) = \{X \subseteq \mathcal{D} \mid \Upsilon^+ \subseteq X \wedge \Upsilon^- \cap X = \emptyset \supset x \in X\}.$$

We call this a *simple reason* ([Doyle 1983a, 1983c]). Simple reasons are the principal attitudes manipulated by reason maintenance systems. (In [Doyle 1979] they are called "support-list justifications.")

It has been popular to present simple reasons in logical encodings. McDermott and Doyle's [1980] nonmonotonic logic encodes each simple reason

$$\{a_1, \ldots, a_i\} \setminus\setminus \{b_1, \ldots, b_j\} \mid\!\!\vdash c$$

as a formula

$$a_1 \wedge \ldots \wedge a_i \wedge M\neg b_1 \wedge \ldots \wedge M\neg b_j \supset c$$

of a logic including a modality M interpreted as "consistent." Reiter's [1980] logic of defaults encodes each simple reason as an inference rule

$$a_1 \wedge \ldots \wedge a_i : M\neg b_1, \ldots, M\neg b_j \vdash c$$

again using M to mean "consistent" but this time as part of the meta-language expressing the applicability of the rule. But both of these encodings are flawed, for they employ the complex, difficult-to-compute property of consistency to encode the simple properties of presence and absence of attitudes. Consistency has nothing to do with the meanings of simple reasons, and more natural logical encodings do not refer to it at all (see [Doyle 1983a,c]). The left-over logical component of nonmonotonic logic and default logic may be better thought of as a sort of "autoepistemic" logic (as Moore [1983] calls it).

3.2 Constitutive logics of mental states

For most agents of interest, we may associate with each state space $\mathcal{I}$ describing the agent a logic of mental elements (or psycho-logic) so that each state in $\mathcal{I}$ corresponds to a deductively closed and consistent set of mental elements in $\mathcal{I}$'s logic. These abstract logics are called *information systems*. Although information systems may be viewed simply as theoretical devices for describing the agent's states, they are actually much more important than that, for many of the sorts of automatic limited deduction and automatic limited consistency checking common in artificial intelligence systems may be described directly in terms of information systems. Automatic reasoning specified by constitutive intentions may go beyond the logic of the information system, and to describe such state spaces we extend the notion of information system to that of *satisfaction system* by incorporating the notion of abstract self-specification.

3.2.1 Information systems

Following Scott [1982], an information system Σ is defined by three things: a set $\mathcal{D}$ of data objects (the "finite" or initial data objects), a set $\mathcal{C}$ of finite subsets of $\mathcal{D}$ (the "consistent" finite subsets), and a relation $\vdash$ on $\mathcal{C} \times \mathcal{D}$ (the "entailment" relation), where if $X \subseteq \mathcal{D}$ and $y \in \mathcal{D}$, we write $X \vdash y$ instead of $(X, y) \in \vdash$. The basic idea is to use these notions to define a data type or *domain* by viewing each individual data object as a "proposition" about domain elements, and each set of data objects as a partial description of some domain element, with bigger sets representing better descriptions. When descriptions contain enough "propositions," the sets of data objects characterize (possibly partial) domain elements, and so we may identify the elements of the domain with these sets of data objects.

The formal notions of consistency and entailment are given some substance by the following axioms. For each $x, y \in \mathcal{D}$ and $X, Y \subseteq \mathcal{D}$, $\mathcal{C}$ satisfies

1. If $X \subseteq Y \in \mathcal{C}$, then $X \in \mathcal{C}$,
2. If $y \in \mathcal{D}$, then $\{y\} \in \mathcal{C}$, and
3. If $X \vdash y$, then $X \cup \{y\} \in \mathcal{C}$.

That is, subsets of consistent sets are consistent; each data object is itself consistent; and addition of entailed elements preserves consistency. The entailment relation $\vdash$ satisfies

1. If $x \in X$, then $X \vdash x$, and

2. If $Y \vdash x$ for all $x \in X$, and $X \vdash y$, then $Y \vdash y$.

That is, consistent sets entail their own members, and entailment is transitive. This is clearer if we extend the notation of entailment in the natural way to say that $X \vdash Y$ iff $X \vdash y$ for each $y \in Y$, in which case the last condition can be rewritten as $X \vdash Z$ whenever $X \vdash Y$ and $Y \vdash Z$.

We say that a set $X \subseteq \mathcal{D}$ is *consistent* if each finite subset $Y \subseteq X$ is consistent according to $\mathcal{C}$. We say that X is *closed* iff $x \in X$ whenever $Y \subseteq X$ and $Y \vdash x$. Putting these things together, the *elements* of the information system, the set of which is written $|\Sigma|$, are the closed and consistent subsets of $\mathcal{D}$. Some of these we view naturally as the partial (or incomplete) elements of the domain, such as the minimal element

$$\bot = \{x \in \mathcal{D} \mid \emptyset \vdash x\}.$$

The *total* (or complete) elements are those elements maximal under set inclusion, that is, elements $X \in |\Sigma|$ such that $Y = X$ if $Y \in |\Sigma|$ and $X \subseteq Y$. We write $\lceil \Sigma \rceil$ to mean the set of total elements of Σ. We say that X *approximates* Y whenever $X \subseteq Y$. Every element in a domain is the limit of its finite approximations, and there is a rich theory of approximation that forms the basis of the theory of computation over domains (see [Scott 1982]). Finally, if X is consistent we define $\theta(X)$ to be the *closure* of X, the least closed superset of X, or in terms of approximations, the least element of $|\Sigma|$ approximated by X. θ thus corresponds to the usual operator Th of deductive closure in ordinary logic.

The view of information systems in which data objects are "propositions" that describe domain elements thus views each data object $x \in \mathcal{D}$ as meaning the proposition $\{X \mid x \in X \wedge X \in |\Sigma|\}$. This view is also natural in that alternatively we can consider the elementary notions of consistency and entailment as stemming from a propositional meaning function. Specifically, every function $[\![\,]\!] : \mathcal{D} \to \mathcal{L}$ that interprets each object $x \in \mathcal{D}$ as an element $[\![x]\!]$ of a complete lattice $\mathcal{L}$ yields natural notions of consistency and entailment. $\mathbf{P}\mathcal{W}$ and $\mathbf{P}\mathcal{D}$, with intersection as meet and $\emptyset$ as $\bot$, are such lattices. If $X \subseteq \mathcal{D}$ we define the

meaning of X to be the meet of the meanings of its elements, that is

$$[\![X]\!] = \bigwedge_{x \in X} [\![x]\!].$$

The consistent sets are those $X \subseteq \mathcal{D}$ such that $[\![X]\!] \neq \bot$, and we say that $X \vdash Y$ iff $[\![X]\!] \sqsubseteq [\![Y]\!]$, so that X is closed iff $Y \subseteq X$ whenever $[\![X]\!] \sqsubseteq [\![Y]\!]$. Under the assumption that $[\![x]\!] \neq \bot$ for every $x \in \mathcal{D}$, the meaning function determines an information system.

Our first constitutive assumption about the agent was that $\mathcal{D} = \mathbf{B} \cup \mathbf{D} \cup \mathbf{I}$, and our second was that each state in $\mathcal{I}$ is satisfying. Our third constitutive assumption is that there is some information system Σ over this $\mathcal{D}$ such that each state of the agent is an element of the domain defined by Σ. That is, $\mathcal{I} \subseteq |\Sigma|$, and each state is closed and consistent with respect to Σ.

The most common sorts of agent states in artificial intelligence are composed of data structures and propositional representations patently amenable to the information system viewpoint. Of course, a single state-space may admit different representations in terms of information systems, each with different notions of data objects, consistency, and entailment. In some cases states may be described exactly as the elements or total elements of some Σ, so that $\mathcal{I} = |\Sigma|$ or $\mathcal{I} = \lceil \Sigma \rceil$, but it is an open question whether this is possible for every state space of psychological interest. Ideally, the specification of the structure of an agent should be indifferent to the choice of information system used to discuss it. (We pursue the question of invariant representations in a subsequent paper.)

Many notions of entailment and consistency fit within the framework of information systems. It should be clear that the ordinary logical notions of consistency and entailment satisfy these axioms. Alternatively, $\mathcal{C}$ need not be taken to be the usual notion of consistency, but may instead capture only lack of overt inconsistency, as in

$$\mathcal{C} = \{X \subseteq \mathcal{D} \mid X \text{ finite } \wedge\ \neg\exists x[x \in X \wedge \neg x \in X]\}.$$

Similarly, $\vdash$ need not be taken to be the usual notion of entailment, but might only capture propositional deduction instead of first-order deduction, or only Modus Ponens ($X \vdash y$ iff either $y \in X$ or there is some $x \in \mathcal{D}$ such that $x \in X$ and $x \supset y \in X$), or only ground instantiation ($X \vdash y$ iff either $y \in X$ or y is a ground instance of some $y' \in X$), or entailment in modal, relevance, or probabilistic logics. The minimal notion of consistency is the vacuous restriction

$$\mathcal{C} = \{X \subseteq \mathcal{D} \mid X \text{ finite}\}$$

in which all sets are consistent. The minimal notion of entailment is pure containment ($X \vdash y$ iff $y \in X$), a relation lacking all nontrivial deduction. When both $\mathcal{C}$ and $\vdash$ are minimal, every set is closed and consistent, so $|\Sigma| = \mathbf{P}\mathcal{D}$.

We can always choose $\vdash$ so that some set of attitudes is present in every state as an unchangeable background to the agent's reasoning. These "axioms" need not just be tautologies, but substantial attitudes as well. In each information system, the fixed background is just the least element $\perp = \{x \mid \emptyset \vdash x\}$. For example, in the minimal information system, $\perp = \emptyset$, so there are no background attitudes determined by the information system. (Since $\mathcal{I}$ need not exhaust $|\Sigma|$, the states in $\mathcal{I}$ may still have non-empty intersection, and hence an unchanging background, even if $\perp = \emptyset$.)

The logic of states need not be confined to logics of belief, but may also specify logics of other beliefs, such as closure and consistency conditions on desires and intentions, or on likelihoods and preferences. Many logics proposed in philosophical logic, such as deontic logics and logics of imperatives might also be cast in these terms. These are useful in expressing the more subtle consistency and completeness conditions mentioned but not pursued in section 2.1.3.

3.2.2 Satisfaction systems

Just as the abstraction of information systems permits discussion of logics of mental states even when the elements of mental states are not just beliefs, we may abstract the notion of constitutive intention away from our assumed psychology of beliefs, desires, and intentions and their internal and external meanings. In the notion of satisfaction system, all we retain is the assumption that some elements of mental states have constitutive meaning. We define a satisfaction system to be an information system $\Sigma = (\mathcal{D}, \mathcal{C}, \vdash)$ together with a meaning function $[\![]\!] : \mathcal{D} \to \mathbf{PP}\mathcal{D}$, the idea being that the meaning $[\![x]\!]$ of an element x is the set of possible states (worlds) that satisfy the constitutive intent, if any, of the element, and that satisfying worlds are closed and consistent sets that satisfy the constitutive import of each of the elements they contain. (If $\mathbf{I}^\star$ varies with the agent's state, then these additional propositional interpretations of data objects will depend on the state as well.) If x has no constitutive import, then it places no restrictions on possible states containing it, so $[\![x]\!] = \mathbf{P}\mathcal{D}$. We extend $[\![]\!]$ to

subsets of $\mathcal{D}$ by defining

$$[\![X]\!] = \bigcap_{x \in X} [\![x]\!]$$

for each $X \subseteq \mathcal{D}$. We define the set $\|\Sigma\|$ of *satisfying* domain elements to be

$$\|\Sigma\| = \{X \mid X \in |\Sigma| \wedge X \in [\![X]\!]\}.$$

Clearly, $\|\Sigma\| \subseteq |\Sigma|$. Note that the definition permits unsatisfiable elements (elements x such that $[\![x]\!] = \emptyset$) which may not appear in any satisfying state.

We assume that there is a satisfaction system Σ over $\mathcal{D} = \mathbf{B} \cup \mathbf{D} \cup \mathbf{I}$ such that $\mathcal{I} = \|\Sigma\|$, thus refining our earlier assumption that $\mathcal{I} \subseteq |\Sigma|$. This assumption may be satisfied trivially if $\perp$ is nonempty and we choose $[\![x]\!] = \mathcal{I}$ for each $x \in \mathcal{D}$, since then $\mathcal{I} = \|\Sigma\|$. The more interesting question, for which we supply no answer, is whether for each satisfaction system Σ there is an information system Σ' such that $\|\Sigma\| = |\Sigma'|$ or $\|\Sigma\| = \lceil \Sigma' \rceil$.[2]

As seen earlier, propositional meaning functions give rise to information systems in quite natural ways, so we may consider the case of the meaning functions $[\![\,]\!]$ appearing in satisfaction systems. The main observation is that even if we stipulate that $[\![x]\!] \neq \emptyset$ for every $x \in \mathcal{D}$, the information system arising from $[\![\,]\!]$ alone is not really of interest. The central idea of satisfaction systems is the idea of self-satisfying states, and self-satisfaction, as opposed to satisfaction, is a notion absent from the information system framework. For example, we might derive an information system $\Sigma^* = (\mathcal{D}, \mathcal{C}^*, \vdash^*)$ from $\Sigma = (\mathcal{D}, \mathcal{C}, \vdash, [\![\,]\!])$ by defining $\mathcal{C}^* = \{X \in \mathcal{C} \mid [\![X]\!] \neq \emptyset\}$, and $X \vdash^* y$ iff $X \in \mathcal{C}^*$ and $X \vdash y$. It is easily checked that

$$|\Sigma^*| = \{X \mid X \in |\Sigma| \wedge [\![X]\!] \neq \emptyset\}.$$

[2]The notion of satisfaction system not only generalizes the notion of constitutive intention to other psychological organizations, it also recasts the notion of admissible state semantics of [Doyle 1983a,e] in simpler terms. In [Doyle 1983a], elements were allowed constitutive meaning and states were satisfying or not just as in the present treatment. The main difference in formulation concerned the background logic, which was not discussed explicitly, but introduced only through a restriction set $\mathcal{R} \subseteq \mathbf{P}\mathcal{D}$, with states of the agent required to be satisfying elements of $\mathcal{R}$. In the present setting, we can better understand the nature of the restriction set by assuming that for most applications $\mathcal{R} = |\Sigma|$ for some information system Σ. (The treatment in [Doyle 1983a] is more general in that a specific $\mathcal{R}$ may not be definable as a domain over $\mathcal{D}$, but this added generality may not be interesting.) This assumption satisfies both of the explicit motivations for choices of $\mathcal{R}$ given in [Doyle 1983a], namely to rule out the empty state, and to require consistency and closure of states. Consistency and closure, of course, are what information systems are about. Nonemptiness of states comes as a consequence of this, if we choose the notion of entailment so that the empty set is not closed.

Since $[\![X]\!] \neq \emptyset$ if $X \in [\![X]\!]$, we see that $\|\Sigma\| \subseteq |\Sigma^*|$, that in fact,

$$\|\Sigma\| = \{X \mid X \in |\Sigma^*| \wedge X \in [\![X]\!]\}.$$

But in general $X \notin [\![X]\!]$, hence $\|\Sigma\| \neq |\Sigma^*|$, so $|\Sigma^*|$ is not of interest.

3.3 Laws of thought and legal actions

While some constitutive intentions may place restrictions on the legal states occupied by the agent, others may restrict the sorts of changes suffered or actions performed by the agent. That is, in addition to actions determined by executing a single intention selected from the agent's plans, we allow actions taken automatically in order to satisfy constitutive intentions about actions. We say that a state change $S_t \mapsto S_{t+1}$ is *satisfying* if it satisfies each of the constitutive intentions contained in S_t. (To express this precisely requires the larger view that intentions are about sets of possible histories.)

Constitutive intentions about actions may be used to capture the dynamic notion of indicative conditionals, as rules for revising one's attitudes upon learning new information, and to express the transition rules for simple sorts of parallel machines, such as cellular automata, where constitutive intentions stipulate the state changes of each cell as a function of its current state and the current states of neighboring cells.

3.3.1 Constitutive priorities

Constitutive intentions about actions add another consistency requirement beyond those imposed by constitutive intentions about states and the constitutive logic, for no satisfying actions are possible if the current set of constitutive intentions specify inconsistent changes. That is, satisfying state changes exist only if the constitutive intentions about actions are *consistent*, that is, if the set

$$\iota^*(S_t) = \bigcap_{x \in S_t \cap \mathbf{I}^*} \iota(x)$$

is nonempty, so that we may have $S_{t+1} \in \iota^*(S_t)$.

One might think that one sort of inconsistency is as bad as another, but inconsistency of intentions in general is at once more common and less serious than other sorts of inconsistencies, such as of beliefs or preferences. Intentions

are frequently inconsistent solely because priorities are lacking. Lacking priorities to separate the times to which the intentions refer lets them all refer to the same time, at which they are mutually inconsistent. When priorities are added, separating the times to which the intentions refer, they need not be inconsistent.

We may employ this observation to permit constitutive priorities that remove inconsistencies between constitutive intentions about actions by postponing some of the intentions about actions, that is changing their interpretation so that they do not apply to the next state but to some (usually unspecified) later state. Such constitutive priorities thus reduce the set of constitutive intentions applicable in the current state. If the priorities are properly chosen, the reduced set of intentions is satisfiable. Of course, for instantaneous intentions postponement may be equivalent to abandonment, since if conditional the constitutive intentions may no longer be applicable in the next state.

3.3.2 Constitutive preferences

If the agent's constitutive priorities do not suffice to reduce the constitutive intentions about actions to a consistent subset, the agent might reduce the set further by rationally choosing a consistent subset, finding, for example, the maximally preferred subset of constitutive intentions. This sort of rational resolution of inconsistent intentions mirrors the rational selection of accommodations treated earlier. In each of these cases, we may simplify the task of rational selection by assuming that the agent does not use all of its preferences in making the selection, but only a subset chosen so that these cases of decision-making are simple and tractable. We may think of these preferences as *constitutive* preferences of the agent. That is, we identify a subset $\mathbf{D}^{\star} \subseteq \mathbf{D}^{\prime}$ of the agent's preferences about its own states and actions as constitutive, as always or automatically satisfied. We assume that constitutive preferences are of a sufficiently simple nature that we may ensure their consistency through the psycho-logic, that is, require that if $X \subseteq \mathbf{D}^{\star}$ and $X \in \mathcal{C}$, then

$$(\bigcup_{x \in X} \pi^{<}(x))^{\star}$$

is consistent. In this way, the notion of consistency captured in the constitutive logic agrees with the notion of consistency resulting from comparison of meanings. One way of doing this might be to always check consistency of the meaning of a potential constitutive preference with the meanings of the current set before adopting it as a new one.

3.3.3 Conditional actions

Several common ideas in artificial intelligence can be naturally viewed as constitutive intentions about actions. Foremost among these is the common sort of condition-action production rule, which states that if the current state satisfies some property, specific elements should be added and deleted to get the next state, or schematically,

$$\kappa \Longrightarrow \partial^+ \setminus\setminus \partial^-.$$

Sets of such rules are compatible (but not necessarily satisfiable) if the total specified sets of additions and deletions are disjoint, that is, if

$$\bigcup_{x \in V(S)} \partial_x^+ \cap \bigcup_{x \in V(S)} \partial_x^- = \emptyset,$$

where $V(S)$ is the set of applicable conditional actions, that is $V(S) = \{x \in S \cap \mathbf{I}^* \mid S \in \kappa_x\}$.

Production system languages such as OPS provide vocabularies for expressing simpler or richer sorts of conditions. The simplest sort of condition checks for presence and absence of specific elements, which we might write as

$$A \setminus\setminus B \Longrightarrow \partial^+ \setminus\setminus \partial^-,$$

whose positive and negative meanings in state S are respectively ∂^+ and ∂^- if $A \subseteq S \subseteq B^c$, and are $\emptyset$ and $\emptyset$ otherwise.

Constitutive priorities about constitutive intentions also appear in artificial intelligence, the foremost appearance being the "conflict resolution" rules employed in condition-action production systems. These systems are designed with the aim of having exactly one production rule applicable once the set of candidates has been reduced by means of the conflict resolution rules. Typical conflict resolution rules consider the order in which the rules appear in a master list, their generality or specificity, and the recency or immediacy of their applicability, preferences which explicitly refer to the history of the agent.

Chapter 4

Representative self-government

When agents are constructed out of many attitudes from disparate sources, the possibility of inconsistency arises. In the preceding, we have largely ignored this possibility, basing our discussion on definitions holding when the likelihoods, preferences, and priorities of the agent are consistent. But one cannot realistically assume inconsistency away. As a computational problem, it may be very difficult, even impossible, and in any event very expensive to tell whether one's attitudes are consistent. As a practical problem, most artificial intelligence systems contain information supplied by several experts, in the case of expert systems, or by populations of fields in the case of encyclopedias. But different experts have different opinions, even when each expert is individually consistent. If all contributors do not resolve their differences before informing the expert system, the expert system must adjudicate them itself.

While in the long run the agent may work to make its attitudes consistent, in the short run it must live with inconsistency. One way of living with inconsistency is to temporarily impose enough consistency to allow rational action, that is, to choose coherent subsets to act upon. These choices of consistent grounds for action may differ with each action taken. Of course, this approach only produces the instantaneous appearance of consistency, and when the agent is observed over time, its internal inconsistencies manifest themselves in its actions. That is, the agent's actions may appear inconsistent with each other in the sense that the attitudes imputed to the agent by an observer will be inconsistent for different actions. Either the observers will develop very elaborate but self-consistent apologies for the rationality of the agent's actions, or they will decide the agent does not know what it thinks. In this way, the inconsistent attitudes

are reflected in inconsistent behavior.

Inconsistency reduces the aim of rational self-government to the practice of mere self-government. Though the notion of ideal rationality provides a standard for action when the agent is consistent, there are no universally accepted standards for how to act under ambiguity, whether the ambiguity results from incompleteness or inconsistency. Just as there are many forms of government known for human societies, each with its own advantages, disadvantages, proponents and foes, there are many forms of self-government, with similarly various advantages and disadvantages. This is no accident. Conflict, or inconsistency of interests, is the heart of political theory, for when groups are unanimous or compatible, there is no politics, only consensual action. How to act when inconsistent is what political theory is all about, and there is a direct correspondence between political organizations for human government and psychological organizations for individual self-government. The conflict of intuitions about inheritance noted in [Touretzky et al. 1987] is real since every plausible system of government is a possible system for inference, inheritance, and consistency resolution.

4.1 Social agents

Since we may not realistically assume that the agent's attitudes are consistent, we must face up to this difficulty in our theories. The natural way to do this is to assume that the agent is composed of many parts (for example, mental attitudes) which appear or act independently of each other. The task then is to describe how the behavior of the whole agent arises out of the behaviors and interconnections of its parts or subagents. We call this the *social* view of agent structure and behavior.

The social agent view appears in psychology as notions of mental organs or faculties, most explicitly in Gazzaniga's [1985] "social brain," Minsky's [1986] "society of mind," and in the modularity hypotheses of Fodor [1983]. These authors view the mind as composed of many disparate organs, faculties, and subagents. Gazzaniga, for example, does not try very hard to motivate this hypothesis in terms of evolutionary biology or computational complexity, but instead focuses on explaining the observed behavior of humans in these terms. In his view, the function of beliefs is to provide a self-image, a theory for the agent to use in understanding its own actions. This theory need not be accurate.

Indeed, it cannot be if it is merely a coherent face put on an underlying conflict of attitudes. Sometimes (perhaps usually) when beliefs conflict with actions, the beliefs are changed (Festinger [1957] calls this "cognitive dissonance"), and sometimes actions are changed. Internal conflicts and imperfect self-knowledge also appear in theories of disappointment, self-commitment, and self-deception (see [Hirschman 1982], [Scitovsky 1976], [Elster 1979, 1983], [de Sousa 1971], [Schelling 1984a], [Maital 1982], and [Kydland and Prescott 1977].)

More obvious cases of social agents are studied in sociology, economics, and political theory, where they appear as groups, corporations, parties, organizations, and agencies. Where psychologists consider minds composed of societies, social theorists treat societies composed of minds, in which the attitudes and behaviors of the group derive from the organization, attitudes, and behavior of its members (see [Pareto 1927], [Berger and Luckmann 1966], [Mundell 1968], and [Mueller 1979]).

The social agent viewpoint comes naturally to artificial intelligence, which early on designed agents organized into collections of separate modules, processors, databases, frames, contexts, K-lines, local problem-solving memories, hierarchical descriptions, and semantic networks (see [Minsky 1986], [Abelson and Sussman 1985]). The social agent view has recently received added attention due to increasing interest in the notion of parallelism. Parallelism is frequently proposed as a means for speeding up computations, but if simultaneous activities are to be coordinated—that is, yield consistent results—then substantial speedups from parallelism are not always possible. That is, some tasks are inherently sequential in that there is no way to decompose them into many simultaneous consistent subtasks. In these cases, seriality is the only way of guaranteeing consistency. But if we permit agents to be inconsistent, these restrictions may not apply, and we might as well employ naive parallelism, even if it makes the agent a bit more inconsistent.

In this chapter we wish to describe the composition and organization of social agents, construed broadly enough to encompass the ideas of artificial intelligence, psychology, and the social sciences. Thus we will view as social agents not just inconsistent individual agents or persons, but also larger bodies (groups, societies, agencies, corporations) and smaller ones (mental faculties and organs). The internal organization of social agents will be described by structured sets of parts or members, and the external appearance of these parts described by the attitudes of the agent and how these attitudes relate to the states of its members, the agents of which it is a member, and other parts of the agent's

environment.

4.1.1 Associations

We use the neutral term *body* to refer to all sorts of social agents, at all levels. The largest unit of organization is the *universal body* Λ, which we assume encompasses all other bodies. We write Ω to mean the set of all possible proper bodies, called the *universe* of bodies. Λ is an improper body, in that we assume that $\Lambda \notin \Omega$, but we define the *closed* universe $\bar{\Omega}$ to be the set of all possible bodies, proper or improper, namely $\bar{\Omega} = \Omega \cup \{\Lambda\}$.

Possible worlds over the closed universe $\bar{\Omega}$ associate organizational and informational structures with each body in $\bar{\Omega}$. We define an *association* $\phi = (\Omega_\phi, m_\phi, \mathcal{I}_\phi, S_\phi)$ to be a collection of mappings over $\bar{\Omega}$, such that for each $A \in \bar{\Omega}$,

1. $\Omega_\phi(A)$ is a subset of Ω, called the *local universe* of A in ϕ,
2. $m_\phi(A) \subseteq \Omega_\phi(A)$ is the set of *members* of A in ϕ,
3. $\mathcal{I}_\phi(A)$ is a set, called the *state space* of A in ϕ, and
4. $S_\phi(A) \in \mathcal{I}_\phi(A)$ is the *state* of A in ϕ.

When the association under discussion is understood, we drop the subscripts and write $\Omega(A)$, $\mathcal{I}(A)$, $m(A)$, and $S(A)$.

We write Φ to mean the collection of all possible associations. Φ is not necessarily a set, since we have not restricted the selection of state spaces in any way, but that will not matter here. In the social agent setting, the set of possible worlds $\mathcal{W}$ is a subset of Φ, propositions are subsets $P \subseteq \mathcal{W} \subseteq \Phi$, and instantaneous states of the world are just associations $\phi \in \mathcal{W}$.

Let $\phi \in \Phi$. We define $m^*(A)$, the *population* of A, to be the set of members of A, members of members of A, and so on. Formally, we define $m^*(A)$ to be the least mapping such that

$$m^*(A) = m(A) \cup \bigcup_{B \in m(A)} m^*(B)$$

for each $A \in \bar{\Omega}$. If we view the graph of m as a relation over $\bar{\Omega} \times \bar{\Omega}$, m^* is just the transitive closure of m. We write $\bar{m}(A)$ and $\bar{m}^*(A)$ to mean, respectively, $m(A) \cup \{A\}$ and $m^*(A) \cup \{A\}$. We define the *global population* ω to be the

population of Λ, that is, $\omega = m^*(\Lambda)$, and write $\bar{\omega}$ to mean $\omega \cup \{\Lambda\}$. We say that bodies *exist* in an association just in case they are in the global population of the association. If $A \in \bar{\omega}$, we define the *environment* A^e of A to be the set of all other existing bodies, that is, $\bar{\omega} \setminus \{A\}$. We define the *proper* environment A^{*e} to be the environment of A's population, that is, $A^{*e} = A^e \setminus m^*(A)$.

If $A, B \in \bar{\Omega}$, we write $A \prec B$ if $A \in \bar{m}^*(B)$. Clearly, the relation $\prec$ quasi-orders $\bar{\Omega}$. We say that a body A is *regular* if $A \notin m^*(A)$, and that ϕ is regular if every body is. If $m(A) = \emptyset$, then $B \prec A$ implies $B = A$ and we say that A is *atomic*. If $\bar{m}^*(A) \cap \bar{m}^*(B) = \emptyset$, we say A and B are *separate*. It is often natural to suppose that bodies communicate or constrain each other's states only through shared members or other incidence of populations.

If $m_\phi(A) = m_{\phi'}(A)$ for all configurations ϕ, ϕ' of a history, we say that A is *rigid.*

The single agent discussed previously may be viewed either as the case $\Omega = \emptyset$ and $\mathcal{I}_\Lambda = \mathcal{I}$ if there is no environment, or as $\Omega = \{A, E\}$, where A is the agent, E its environment, $\Omega(A) = \emptyset$, and $\mathcal{I}(A) = \mathcal{I}$.

4.1.2 Attitudes

For the cases of interest here, it is natural to interpret the states of bodies as sets of attitudes of the body, just as we interpreted states of the unitary agent earlier. As before, we may describe each state space $\mathcal{I}_\phi(A)$ with an information system $\Sigma_\phi(A) = (\mathcal{D}_\phi(A), \mathcal{C}_\phi(A), \vdash_{\phi,A})$ which we call a *framing* of states of A in ϕ if $\mathcal{I}_\phi(A) \subseteq |\Sigma_\phi(A)|$.

If there is a universal set of attitudes $\mathcal{D}$ and set of framings such that $\mathcal{D}_\phi(A) \subseteq \mathcal{D}$ for each $A \in \Omega$, we say the framings are *uniform.* Each uniform set of framings induces a partial order $\sqsubseteq$ on bodies, called the *view* order, such that $A \sqsubseteq B$ iff $S(A) \subseteq S(B)$ in the framings. While different bodies may be atomic among $\bar{\omega}$, there is only one atomic set of attitudes, namely the empty set. This may occur when a body has only one possible state. In general, when a body's state is constant in all associations in a history, we say the body's attitudes are *rigid.*

When states are sets of attitudes, we may view the instantaneous structure of bodies as a uniform set of elements or contents $c(A) = m(A) \cup S(A)$. When attitudes are drawn from a universal set $\mathcal{D}$, we may think of attitudes as atomic bodies, with no attitudes of their own, and $c(A)$ as an expanded notion of membership. In this case, description of the social organization structure simplifies

to a universe $\Omega' = \Omega \cup \mathcal{D}$ and a membership function $m' = c$. This allows us to combine the local information systems describing the framings of each body into a global information system $\Sigma_\phi = (\mathcal{D}_\phi, \mathcal{C}_\phi, \vdash_\phi)$ by defining

1. $\mathcal{D}_\phi = \bigcup_{A \in \bar{\Omega}} \{A\} \times \mathcal{D}_\phi(A)$,
2. $\mathbf{i}_A(X) = \{x \in \mathcal{D}_\phi \mid (A, x) \in X\}$ for each $A \in \bar{\Omega}$ and $X \subseteq \mathcal{D}_\phi$,
3. $\mathcal{C}_\phi = \{X \subseteq \mathcal{D}_\phi \mid X \text{ finite } \wedge\ \forall A \in \bar{\Omega} \quad \mathbf{i}_A(X) \in \mathcal{C}_\phi(A)\}$, and
4. $X \vdash_\phi Y$ iff $\forall A \in \bar{\Omega} \quad \mathbf{i}_A(X) \vdash_{\phi,A} \mathbf{i}_A(Y)$.

This information system combines the local systems without introducing any additional structure. If we wish to add further consistency and entailment conditions, we might allow any Σ_ϕ such that $\mathcal{D}_\phi$ and $\mathbf{i}_A$ are defined as above, with $\mathcal{C}_\phi$ and $\vdash_\phi$ compatible with the local conditions in the sense that for each $A \in \bar{\Omega}$, $\mathcal{C}_\phi(A) = \{\mathbf{i}_A(X) \mid X \in \mathcal{C}_\phi\}$ and $\mathbf{i}_A(X) \vdash_{\phi,A} \mathbf{i}_A(Y)$ iff $X \vdash_\phi Y$.

Each of the notions of meanings, laws, actions, and accommodations previously developed in the context of single agents extends to the case of social agents. Where before propositions were sets of possible worlds, they are now sets of possible associations, so that attitudes that we earlier viewed as self-referential laws of thought are now better viewed as attitudes about the social agent's overall social organization. The earlier distinction between internal and external meanings of attitudes carries over and generalizes to the notion of local meanings, that is, the meaning of one body's attitudes toward another body. For example, it is natural to consider intentions conditional on and concerning an agent's membership and non-membership, and intentions conditional on and concerning the agent's member's states and other aspects of its environment. Consider intentions like "any member of one of my members should also be one of my members," "A and B should not both be members at the same time," and "if each of my members believes something, so should I."

The notion of laws of thought generalizes to local laws of thought, so that the attitudes of each body determine a set of *legal* states for the universe, but legal only with respect to that body. Since a body's members are, strictly speaking, part of its environment, states of the members may fail to satisfy the body's attitudes even if the body's own state does satisfy its attitudes. That is, bodies may be mistaken in their beliefs about their members, and members may be in states the body does not desire or intend. Members that do not satisfy a body's intentions toward them are said to be in *illegal* states with respect to the body.

As before, we may think of some intentions as constitutional in the sense that they define the actual legal states of an agent. But the class of laws that may be taken as constitutional is considerably broader than before. Previously, whether an intention was constitutive or not was not up to the agent to decide. But within social agents, we may consider laws adopted by one body that determine what intentions are constitutive for another body at that time, or for itself in the future. In this setting, constitutions, charters, laws, regulation, rules, and organizational structures are all much alike in general effect, differing mainly in their sources (who or what makes them) and in their fixity (how hard they are for the body itself to change, or for their sources to change). If $\mathbf{I}^*_t(A)$ is the set of constitutive intentions of A at t, then $\phi = W_t$ is a *legal* association if ϕ satisfies every constitutive intention of every body. In the dynamic case, these laws concern not only the changes of attitudes of the various members, as before, but also changes in membership. When these involve changes in the global population ω, we may think of them as cases of incorporation and disincorporation, birth and death, or creation and destruction of agents. Such laws are familiar as stated in human societies, where corporations are just legal persons, persons or bodies created by and regulated by laws. The most familiar examples in computation are actions that allocate and deallocate memory elements in a heap.

In the social agent case, the notions of action and accommodation generalize to local action and accommodation. We allow that several bodies may take actions independently, rather than there being a single effector of changes. These local actions may change, for example, only the state of the body itself, or only a few other bodies with which it communicates. Accommodation to changes may also occur locally, with different versions of accommodation (for example, different constitutive preferences) employed in different bodies.

4.2 Representation

We may think of an association as describing a system of representation in the logical sense, that is, as an interpretation or model in logic. Specifically, we may view associations as (partial) interpretations of a language of typed variables, viewing each body $A \in \Omega$ as a typed variable. In this view, associations interpret the types of bodies as their constitutions $\mathcal{I}_\phi$, and bodies themselves as values S_ϕ within these types. Generally speaking, we may embed any system of typed variables and their interpretation in a universe of bodies and an association.

One important example of this is the notion of attitudinal states. Earlier we remarked that one might treat attitudes as atomic bodies with no attitudes of their own. But this does not mean that attitudes must have trivial state-spaces. Instead, we may think of meanings of attitudes as states of attitudes. Here we assign $\mathcal{I}_\phi(A) = \mathcal{P}$ or $\mathcal{I}_\phi(A) = \mathcal{Q}$ to each attitude A, so that $S_\phi(A)$ may be defined to be the current meaning of A. This permits us to express both the attitudinal and semantic structures of states within a single notion, that of association.

But systems of representation involve more than just the notions of sign and object. They also involve notions of compositionality, or how the meaning of a complex relates to the meanings of its parts. Questions of compositionality are central to both logical and political theories of representation. In the case of logical representations, this means asking how the interpretation or truth value of a set of sentences depend on the interpretations or truth values of the individual sentences and their syntactical components. Most logical connectives and operators (the exceptions being the modalities) are meaning-functional, that is, they determine the meaning of a sentence as a fixed function of only the meanings of the syntactical components. In the case of political economy, compositionality concerns how the attitudes and decisions of a group or official depend on or represent the attitudes and decisions of the members of the group or constituents of the official. Some normative theories of political representation strive for meaning-functionality, ruling out decision rules that depend on things other than the attitudes of members.

In much of what is treated as representation in artificial intelligence systems, constitutive psycho-logics and constitutive intentions relate the states or contents of one subsystem or body to the states or contents of others. That is, the laws of thought or organization describe the representation relation between the subsystem and its constituents. Inheritance networks and constraint systems are good examples. In inheritance systems, individual concepts are represented as bodies, with features or properties represented as the attitudes of the concepts. Laws of organization called inheritance links specify that the attitudes of one concept include or are otherwise derived from the attitudes of other concepts, for example its superiors or superconcepts. (See [Doyle 1980] and [Doyle 1983d] for treatments of inheritance along these lines.) In constraint systems, concepts are represented as bodies as in inheritance systems, but here the laws of organization are local to each concept or constraint. Each constraint may have several parts—subordinate bodies that may be shared with other constraints—and the constraint's rules of organization specify how the states of these parts are derived

from each other. (See [Sussman and Steele 1980] and [Doyle 1983d] for more on this.)

4.2.1 Rational representation

Many special problems of representation result when one restricts the sorts of bodies and constitutions to be of special forms. One of the most interesting cases is that of rational representation, in which the states of all bodies are assumed to be rational. This is the case studied in political economy, which asks how members of a committee, bureaucracy, market, or electorate control the actions of the committee, agency, exchange, or government. The standard theories assume the members to be ideally rational agents, and aim for the composite body to be rational as well. Rational representation is thus exactly the notion of interest in designing agents whose actions should be rational even though the agent's attitudes may be inconsistent. In this setting, the agent's attitudes are the conflicting individuals, and the agent's selection of attitudes on which to act is the set of group attitudes. We have already defined the agent's attitudes so that they are very simple rational agents. Specifically, each attitude can be viewed as an agent whose attitudes are described by its meaning. Since we assumed these meanings are quasi-orders, we already have that the attitudes of attitude-agents are consistent, though incomplete. Thus the agent's task in acting can be viewed as formally identical to the problem of group decision for these attitude-agents. (See [Doyle 1988c] for an application of this idea to default theories. [Nowakowska 1973] makes a similar point. See also [Levi 1986].)

In the context of rational representation, the laws relating member states to group states are called decision rules or aggregation rules. A large literature is devoted to the study of the properties or efficacy of different sorts of decision rules. Some of the most studied sorts of decision rules are those satisfying compositionality criteria. For example, the Pareto condition is that the group should agree with the members whenever the members are unanimous, and the non-dictatorship condition is that the group's attitudes should not reduce to a function of the attitudes of a single member of the group.

The first problem of rational representation is to determine when rational representations exist. Rational representations always exist when no conditions are placed on the decision rules involved. But famous results show that no decision rule satisfying several general but mild conditions can guarantee that the group attitudes are rational. Other results show that if the member attitudes

are sufficiently coherent, for example if their statistics are single peaked, then some simple decision rules like voting give rational group attitudes. And in the case in which attitudes are complete enough so that the utility functions are smooth, general equilibrium theory shows the existence of price systems which represent the preferences of traders in markets. See [Arrow 1963], [Black 1963], [Buchanan and Tullock 1962], [Mueller 1979], [Debreu 1959], and [Arrow and Hahn 1971].

The second problem of rational representation is that even if rational representations exist, they need not be unique. In general there are many ways of resolving conflicts, some ways satisfying one set of members, and other ways satisfying other members. This non-uniqueness is reflected in one of the central notions of group choice theory, namely Pareto optimality. A decision (or set of group attitudes) X is Pareto optimal if any other selection fails to satisfy some preference satisfied by X. Thus if one changes the decision to satisfy some preference unsatisfied by X, one does so at the cost of failing to satisfy some other preference that was satisfied by X. Pareto optimality alone is a very weak condition, however, and most systems of government go far beyond it.

Rational representation plays a big role in artificial intelligence, for we may view some of the conflict resolution rules of production systems, the preferences guiding conservative accommodation, and default rules of frame systems in this light. In the contexts of conservative accommodation and conflict resolution rules, Pareto optimality means making the selection on the basis of some maximal consistent subset of the current preferences. See [Doyle 1985], which interprets a theorem of [Doyle 1983a] along these lines.

4.2.2 Self-representation

One special case of rational representation is rational self-representation. While in rational representation the attitudes of the group represent the attitudes of its members to the external world, so that the agent appears to be a single body with the attitudes of the group, the group body need not have any realization in the world independent of its members. This observation is behind the methodology of individualism in economics, which views collective agents purely as functions of their members, with no member-independent state or existence. But in psychology, it is common to think of the group agent as something more substantial. Psychology tries to view agents as real individuals (humans, after all, have their own bodies), even though they are composed of many conflicting

parts. This assumption simplifies discussion of reflective agents that think about their own states and relations to their environments. (The field of "rational expectations" in economics also treats reflective agents, but without substantiating the group agent.) The simplest way of substantiating the agent's rational selections of attitudes is to introduce a body that, as a member of the agent, contains the rationally selected attitudes and so represents the agent's attitudes to itself as well as to the environment. It is natural to call this special body the agent's self-image.

Formally, we may think of the self-image as a body ME contained in the agent ($ME \in m(A)$), and that the agent's constitutive attitudes describe how the contents $c(ME)$ relate to the general contents $c(A)$ of the agent. In some cases, we may expect that the contents of the agent determine the contents of the self-image, but in other cases we may allow that the agent reorganizes its attitudes to conform with the self-image. In the Freudian vocabulary, we may think of the overall agent attitudes as the id and the self-image as the ego, with the constitutive attitudes, distributed among id and ego or possibly separated into a third body, as forming the superego.

As the Freudian terminology suggests, it is natural to think that there may be a variety of individually embodied self-images, each used in different circumstances. In the Freudian case, one self-image (the ego) is used descriptively, and the other (the superego) is used normatively. In our earlier discussion of deliberation we introduced the notion of decision situation, a compilation of all the attitudes selected as grounds for making a decision. We may easily view each decision situation as a very special self-image, namely the image of one's self with respect to the decision in question. One may also view the idea of temporal or hypothetical contexts in artificial intelligence systems as cases of special-purpose self-images.

When the agent's self-image is used to represent a consistent selection from among the agent's general attitudes, it is a contraction or subset of the agent's full set of attitudes, that is, $c(ME) \subseteq c(A)$. But we may also apply the notion of self-image in cases where the agent's attitudes are consistent but incomplete to represent more complete grounds for action. That is, if there is no action rational given the agent's incomplete attitudes, the agent may represent a completion of its attitudes with a self-image. In this case the self-image is an expansion or superset of the agent's actual attitudes, that is, $c(A) \subseteq c(ME)$. These two cases may be thought of, respectively, as the will to cohere and the will to believe. Of course, the agent may also construct grounds for action that both resolve some

conflicts and fill some incompletenesses, so that the self-image is neither a subset or superset of the agent's attitudes. This is the situation faced in nonmonotonic logic, which resolves some conflicts between defaults by splitting the attitudes into different selections, and at the same time adds some assumptions to each. The same situation appears in theories of inheritance with exceptions [Touretzky 1986]. As these examples show, the existence and uniqueness of rational self-representations is at least as problematic as for plain rational representations.

Chapter 5

Comparative self-government

As was suggested earlier, there is no easy solution to the problem of inconsistency. The complexity of the world makes inconsistent attitudes practically unavoidable, and there is no unique ideal of how to behave though inconsistent. We have identified systems of government as ways of constructing bases for actions: ideally, bases that are rational even though they omit conflicting attitudes and are incomplete with respect to entailment. In artificial intelligence, systems of government are called *architectures*. Going further, we suggest there is also no easy solution to the problem of infeasibility of entailment. Even though there is, in this case, a unique ideal standard, that of deductive closure, allocation of effort towards determining partial sets of entailments is an economic problem, to which there may be several possible rational solutions, each yielding a different set of discovered entailments. The mulplicity of approaches to action and allocation reflect different solutions to the economic problem. In artificial intelligence, the question of how to allocate instantaneous resources of effort, attention, memory, and time is called the *control problem*, and patterns or procedures for making such allocations are called *control structures*.

In general, the agent's degree of rationality reflects the results of its decisions in reasoning, that is, the efficacy of its control decisions about how to reason, or its wisdom in allocating resources. Thus these degrees represent primarily economic limits, not logical limits. That is, the path of reasoning depends on the agent's motivations and choices, not just on logic and the agent's beliefs. All of the agent's attitudes play a role in determining its motivations and choices, and, through these, the agent's mental resources. The agent's resources are always changing anyway through consumption and possibly through changes in

the agent's environment, but the supplies of the most important mental resources are not fixed. Instead, they are what the agent makes them through investment of effort in their improvement or destruction. Hence there is no natural logic of the limits to reasoning, as these limits are not just a matter of the agent's beliefs. Each logic of limited reasoning (such as Levesque's [1984] logic of explicit and implicit belief, or Davis' [1981] logic of obvious inferences) reflects a fixed set of limits, and no such static logic applies to minds possessed of the ultimate resources of intelligence and industry applied over time. Instead one has only degrees of rationality that may vary over time with the agent's actions.

Unfortunately, ideal rationality can not be approximated satisfactorily. Even if one agent is more rational than another, its actions need not be more rational in their results. For example, consider the ideal play in chess approximated by lookahead at most n ply deep. In this case, a better approximation may completely change the results of the previous approximation, even if the results of the shallower search agree with the ideal play. In deliberation, additional steps of reasoning may completely reverse tentative decisions, even if the reversed decision is ultimately the correct one. Thus better approximations usually are only less limited in resources used or attitudes satisfied, not closer in result to the ideal. In economics, the impossibility of approximating ideally rational results by applying increasing resources is called the problem of the "second best," and reflects the fact that partial optima (optima given stipulated boundary conditions) are generally not global optima.

Each architecture or psychological organization involves different sorts of resources. Since each approach suggests different dimensions or ways in which rationality might be limited, we may compare different approaches with respect to the degree of rationality they offer. These comparisons are necessarily incomplete, since there are many dimensions that may be considered separately in each type of government, and different types of government present different sets of dimensions. For example, in the setting of the previous chapters, the agent may be more or less rational with respect to beliefs separately from its rationality with respect to desires or intentions. Thus the ordering of governments by their rationality will be nonlinear, and selections of governments for designed agents will force tradeoffs among desirable qualities. More generally, the intelligence of the agent depends on its state of knowledge as well as its degree of rationality, and comparisons of states of knowledge have even more dimensions than does rationality, for there are many more separate types of knowledge (factual, evaluative, procedural) and subjects of knowledge than dimensions of rationality.

In this chapter, we offer some first steps toward formalizing comparisons of degrees of rationality and states of knowledge as a number of quasi-orders over agents, states, and meanings. These may then be applied to make overall comparisons of architectures, though we do not do so here. In the simplest case, we may see if one system of government dominates another, that is, if it yields actions of greater rationality than another when in the same state. More generally, if we have expectations about the situations to be encountered by the agent and the structure of its knowledge, we may numerically represent the quasi-orders and compare the expected degrees of rationality of the different systems of government, that is, whether one system yields more actions of comparable rationality over time when averaged over possible circumstances.

5.1 States of knowledge

We compare states of knowledge of the agent according to their completeness in each subject, including both internal and external subjects. Rather than simply comparing the sets of attitudes directly, we compare their meanings. That is, we will not consider one agent more knowledgeable than another just because it has numerically more attitudes if the import of those attitudes is the same. Formally, we first define orders on propositions and quasi-orders by saying that $P \leq P'$, for $P, P' \in \mathcal{P}$, iff $P' \subseteq P$ (that is, P' is stronger or more meaningful than P), and that for $M, M' \in \mathcal{Q}$, $M \leq M'$ just in case $M^{<} \subseteq M'^{<}$ and $M'^{\sim} \subseteq M^{\sim}$, so that bigger orders make more strict distinctions and are neutral about fewer things.

With these orders on meanings, we may say that one consistent set of attitudes is greater than another consistent set if the combined meanings of each type of attitude in one set is greater than the combined meanings in the other. Thus if $X, Y \subseteq \mathcal{D}$ are consistent sets of attitudes, we define $X \leq Y$ to mean that

1. $\beta(X \cap \mathbf{B}^a) \leq \beta(Y \cap \mathbf{B}^a)$ in $\mathcal{P}$,
2. $\delta(X \cap \mathbf{D}^a) \leq \delta(Y \cap \mathbf{D}^a)$ in $\mathcal{P}$,
3. $\iota(X \cap \mathbf{I}^a) \leq \iota(Y \cap \mathbf{I}^a)$ in $\mathcal{P}$,
4. $\lambda(X \cap \mathbf{B}^r) \leq \lambda(Y \cap \mathbf{B}^r)$ in $\mathcal{Q}$,
5. $\pi(X \cap \mathbf{D}^r) \leq \pi(Y \cap \mathbf{D}^r)$ in $\mathcal{Q}$, and

6. $\varpi(X \cap \Gamma) \leq \varpi(Y \cap \Gamma)$ in $\mathcal{Q}$.

This means that if $X \leq Y$, an agent with beliefs Y knows at least as much as an agent with beliefs X, even if $X \not\subseteq Y$. When we cannot assume agents are consistent, we say that one set of attitudes represents more knowledge than another if consistent selections from the two can always be made so that selections from one dominate selections from the other. Formally, we say that $X \leq Y$ for any $X, Y \subseteq \mathcal{D}$, if for each consistent set $X' \subseteq X$ there is a consistent set $Y' \subseteq Y$ such that $X' \leq Y'$. This means that the more knowledgable agent may always choose a consistent basis for action that equals or exceeds the choice of the less knowledgable agent.

5.2 Degrees of rationality

Rationality involves both coherence and optimality of actions, so degrees of rationality compare the degrees of coherence and degrees of optimality exhibited by agents.

We may compare the coherence of two agents by comparing the notions of consistency they embody. To do this, we define a relation $\leq$ between information systems $\Sigma = (\mathcal{D}, \mathcal{C}, \vdash)$ and $\Sigma' = (\mathcal{D}', \mathcal{C}', \vdash')$ such that $\Sigma \leq \Sigma'$ if

1. $\mathcal{D} \subseteq \mathcal{D}'$,
2. $X \in \mathcal{C}$ if $X \in \mathcal{C}'$ and $X \subseteq \mathcal{D}$, and
3. $X \vdash' x$ whenever $X \vdash x$ and $X \in \mathcal{C}'$.

It follows that $\leq$ is a partial order on information systems, and that $|\Sigma'| \subseteq |\Sigma|$ if $\Sigma \leq \Sigma'$ and $\mathcal{D} = \mathcal{D}'$. We may use this order on information systems to compare degrees of coherence by using information systems to describe the different logics exhibited by agents at each instant. (Later, we will use this same order to compare the strength of constitutions of agents.) The order on information systems compares at least some aspects of coherence since agents with larger sets of coherent subagents will be more consistent. That is, if we suppose that some bodies are constitutively consistent (as we supposed individual attitudes to be), or that some sets of bodies are constitutively conflict-free, then these suppositions will be reflected in the structure of the consistency notion of the global information system Σ_ϕ. There are other notions of coherence of

importance in reasoning that go beyond simple consistency, for example, single-peakedness of the individual preferences, but we will not pursue these notions here (see [Mueller 1979]).

To compare degrees of optimality, we must think of degrees of rationality not in terms of resources used, but in terms of the grounds or conditions of rationality, in terms of the attitudes accounted for and satisfied by an action. To do this, we must separately compare actions and accommodations with respect to preferences, since both choices of intended action and choices of accommodation involve optimization. Formally, we say that an agent A is more rational than an agent B just in the case that when both are in the same state of knowledge, the consistent sets Y and Y' of preferences used by A in selecting, respectively, the intent and the accommodation, dominate the corresponding sets Z and Z' used by B, that is $Z \leq Y$ and $Z' \leq Y'$. Though it might seem to add little to the definition of one agent being more knowledgable than another with respect to preferences, this definition is different in that it compares the preferences actually used to choose the intent and accommodation, not just some possible selections of preferences.

Comparing the intents of actions compares their progressiveness, while comparing the accommodations of these intents compares their conservativeness. A common confusion occurs when we compare the progressiveness or conservativeness of full actions rather than separately comparing volitions and accommodations. The progressiveness order views all changes as good, while the conservativeness order views all changes as bad, so the more conservative a complete action seems, the less progressive it will be, and vice versa. When we require that intents be optimally progressive and that accommodations be optimally conservative, actions may satisfy both requirements simultaneously and constitute true limited rationality.

It may be, however, that some applications of progressiveness and conservativeness refer not to comparative notions but to absolute ones, with absolute progressive agents being those that make a specified minimum amount of progress with each action, and absolutely conservative agents being those that make no more than a specified maximum amount of change with each action. These absolute notions might best be viewed as reflecting the constitution of the agent, as treated below.

Because they concern how much change the agent makes in acting, degrees of progressiveness and conservativeness are important measures of the subjective speed of the agent, that is, its internal time-scale. Different agents, or different

implementations of the same agent, may appear slower or faster than one another when compared to each other or to the rest of the external world.

5.3 Strength of constitution

Beyond the amount of knowledge and degree of rationality, the level of intelligence of an agent depends also upon the strength of its constitution. We already have most of the means for formally comparing constitutional strength of agents in the comparisons $\Sigma \leq \Sigma'$ on information systems and $X \leq X'$ on states of knowledge. The constitutive logic of agents may be compared through the satisfaction systems describing their state-spaces. By applying the order on intentions to compare constitutive intentions, the partial order on information systems may be extended to one on satisfaction systems by saying that $\Sigma \leq \Sigma'$, for $\Sigma = (\mathcal{D}, \mathcal{C}, \vdash, [\![\,]\!])$ and $\Sigma' = (\mathcal{D}', \mathcal{C}', \vdash', [\![\,]\!]')$, if $\Sigma \leq \Sigma'$ as information systems and if $[\![x]\!]' \leq [\![x]\!]$ in $\mathcal{P}$ for each $x \in \mathcal{D}$. Then if $\Sigma \leq \Sigma'$, we have $\|\Sigma'\| \subseteq \|\Sigma\|$. Similarly, we use the order on states of knowledge to compare the sets of constitutive priorities and preferences used by the agent in the progressive and conservative stages of action.

We may thus consider levels of intelligence to be characterized by constitutive lower bounds on knowledge and degrees of rationality. While even brilliant thinkers may make allocation errors and so take stupid actions, dullards suffer from abnormally low lower bounds on their rationality. It isn't that they can't perform the same reasoning in principle, but that so much more must be made conscious, at enormous costs in attention and resources. Since the bookkeeping needed to keep attention focused must itself be attended to, the result is that the half-wit will find things not uniformly twice as hard, but exponentially (in the complexity of the problem) harder than the full-wit. This is a handicap even extreme diligence will find hard to conquer. One important case of this is the difference between novices and experts. Novices, even if possessed of adequate instructions, must perform every step consciously, and expend much effort in keeping track of their progress and of their understanding of the instructions. For the expert, almost all of this reasoning is automatic, and seems effortless. Normal novices have adequate automatic reasoning powers, but have not yet committed their instructions to these powers. They may be intelligent in other arenas, but in their new subject they are stupid.

Chapter 6

Conscious self-government

One of the lessons implicit in the preceding is that consciousness is an activity, not simply a property enjoyed by an agent. In artificial intelligence and philosophy, it is customary to think of reflection and representation as passive relations between agent and object, not as actions the agent intends or performs. But in making and changing assumptions, in planning, and especially in deliberation, we see that reflection is a process of reasoning and choice that the agent employs routinely in the course of living, so that self-delineation, self-definition, self-evaluation, and self-modification are the very basis of rational self-government.

The curious thing about consciousness, however, is its dispensability, not its necessity. In almost every specific activity necessary to human life, self-consciousness is unnecessary. Jaynes [1976] expands on insights from ethology (see, for example, [Tinbergen 1951]) to illustrate this dispensability in many basic human activities, and current work on expert systems illustrates its dispensability in many non-basic human activities. Recent work on routines by Ullman [1983] and Chapman and Agre [1987] pushes this observation even further.

But if consciousness is dispensable, when is it desirable? Consciousness's major benefit is that it makes the evolution or improvement of its practitioners more rapid. The rapidity of the evolution of human living in historical times has no equal in macroscopic biological evolution, and the rapidity of evolution of human living under capitalism, with its deliberate pursuit of increasing productivity, has few equals in human history. (See [Jaynes 1976], [Festinger 1983], [Drucker 1985], and [Rosenberg and Birdzell 1986] for interesting theories of

the role of self-consciousness in human history.)

Though consciousness has many benefits even when it is not strictly necessary, the benefits of consciousness do not come free. Consciousness misapplied during routine actions interrupts their performance, or at least slows it down. If consciousness becomes an end in itself or turns on itself, paralysis may result. (See [Schön 1983], [Dreyfus and Dreyfus 1985].)

Consciousness has some deeper consequences besides these benefits and costs regarding the efficacy of reasoning. One consequence is that conscious self-government entails personhood. Frankfurt [1971] has developed a theory of personhood in which the criterion of personhood is the possession of attitudes towards one's own attitudes. Such attitudes form the backbone of rational self-government. A more fundamental consequence of conscious self-government is absurdity. Without consciousness, people could not develop feelings of self-insignificance, appreciations of the absurd. (See [Nagel 1979], [Doyle 1980].) Lesser beings never destroy themselves. Conscious agents, in choosing their own evolution, must ask: "By whose values?" The answer is by those they have inherited or adopted, but they may also ask: "Why these and not others?" In contemplating this question, the meaning of existence is questioned, for it is easy to see paths to states in which removal of all values and beliefs is rational. And that is the end.

If when my wife is sleeping
and the baby and Kathleen
are sleeping
and the sun is a flame-white disc
in silken mists
above shining trees,—
if I in my north room
dance naked, grotesquely
before my mirror
waving my shirt round my head
and singing softly to myself:
"I am lonely, lonely.
I was born to be lonely,
I am best so!"

If I admire my arms, my face,
my shoulders, flanks, buttocks
against the yellow drawn shades,—

Who shall say I am not
the happy genius of my household?

William Carlos Williams, *Danse Russe*

Appendix A

Immediate implication and inconsistency

Under one interpretation, the primitive formal notions of information systems, "consistency" and "entailment," correspond naturally to Harman's [1986] primitive (but informal) notions of "immediate inconsistency" and "immediate implication." Harman develops his notions as the only fragments of logic visible in the psychology of ordinary reasoning. When formalized as information systems, these notions provide the logic of the structure of mental states, fragments of logic visible in the mental states themselves.

Immediate implications and inconsistencies can be interpreted formally in either of two ways. In the first way, the sets of immediate implications and inconsistencies are fixed throughout the history of the agent. For $X \subseteq \mathcal{D}$ and $x \in \mathcal{D}$, we write $X \supset_{\mathrm{i}} x$ if the elements of X immediately imply x, and we write $\neg_{\mathrm{i}} X$ if the elements of X are immediately inconsistent.

We can construct an information system from these immediate notions by defining $\mathcal{C}^{\mathrm{i}}$ to be the finite sets not containing any immediately inconsistent subsets, that is, saying $X \in \mathcal{C}^{\mathrm{i}}$ iff

1. X is finite, and

2. there is no $Y \subseteq X$ such that $\neg_{\mathrm{i}} Y$,

and by defining $\vdash^{\mathrm{i}}$ to be provability using Modus Ponens on immediate implications, that is, saying $X \vdash^{\mathrm{i}} e$ iff $X \in \mathcal{C}^{\mathrm{i}}$ and either

1. $e \in X$, or

2. there is a sequence $e_0, \ldots, e_n$ of elements of $\mathcal{D}$ such that $e = e_n$ and for each i, $0 \leq i \leq n$, either

   (a) $e_i \in X$, or

   (b) there is some $Y \subseteq \mathcal{D}$ such that $Y \supset_{\mathrm{i}} e_i$ and for each $y \in Y$ there is some $j < i$ such that $y = e_j$.

It is easy to see that $\mathcal{C}^{\mathrm{i}}$ and $\vdash^{\mathrm{i}}$ form an information system Σ^{i} just in case no single element or immediate implication is inconsistent, that is, just in case there is no $x \in \mathcal{D}$ such that $\neg_{\mathrm{i}}\{x\}$, and just in case $X \cup \{x\} \in \mathcal{C}^{\mathrm{i}}$ whenever $X \supset_{\mathrm{i}} x$.

A very similar construction suffices to extend an initial information system $\Sigma = (\mathcal{D}, \mathcal{C}, \vdash)$ to an information system $\Sigma' = (\mathcal{D}, \mathcal{C}', \vdash')$ that combines $\mathcal{C}$ and $\neg_{\mathrm{i}}$ to get $\mathcal{C}'$, and combines $\vdash$ and $\supset_{\mathrm{i}}$ to get $\vdash'$. It does not matter if some of the immediate implications and inconsistencies redundantly restate some of the entailments of Σ. Note that Σ' extends both Σ and Σ^{i}, that is, $\Sigma \leq \Sigma'$ and $\Sigma^{\mathrm{i}} \leq \Sigma'$.

In either of these cases, each immediate implication or inconsistency directly represents an entailment or consistency condition. In the second interpretation immediate implications and inconsistencies are constitutive intentions contained in states, so that the implicit immediate logic may vary from time to time. In this case we view $X \supset_{\mathrm{i}} x$ and $\neg_{\mathrm{i}} X$ as elements of $\mathcal{D}$ such that

$$[\![X \supset_{\mathrm{i}} x]\!] = \{S \subseteq \mathcal{D} \mid X \subseteq S \supset x \in S\}$$

and

$$[\![\neg_{\mathrm{i}} X]\!] = \{S \subseteq \mathcal{D} \mid X \not\subseteq S\}.$$

Here we may construct an information system $\Sigma(S)$ for each state $S \in \|\Sigma\|$ to represent the instantaneous logic expressed in that state. The construction is as above, only we use only those immediate implications and inconsistencies contained in S.

Appendix B

Temporal and logical nonmonotonicity

The adjective "nonmonotonic" has suffered much careless usage recently in artificial intelligence, and the only thing common to many of its uses is the term "nonmonotonic" itself. In fact, two principal ideas stand out among these uses: namely, that attitudes are gained and lost over time, that reasoning is nonmonotonic—this we call *temporal* nonmonotonicity—and that unsound assumptions can be the deliberate product of sound reasoning, incomplete information, and a "will to believe"—which we call *logical* nonmonotonicity. Indeed, much of the confusion reigning about the subject stems from a confusion between these two sorts of nonmonotonicity, and between logical nonmonotonicity and nonmonotonic logic.

Let us differentiate these uses in precise formal terms. In mathematics, the terms monotonic and nonmonotonic (or monotone and nonmonotone) refer to properties of functions between ordered sets, so to use these terms with precision in describing a reasoning agent, we must identify specific functions with ordered domains and codomains to which we may attribute these properties. When we view states through the lens of information systems as closed sets of mental attitudes, we can distinguish two functions between ordered sets. Let $\mathcal{T}$ be the set of temporal instants of a history ordered by increasing time. Consider the state space $\mathcal{I}$ as a subset of $\mathbf{P}\mathcal{D}$ ordered by set inclusion, so that states with additional elements are considered bigger. We may then view each history of the agent as a function

$$S : \mathcal{T} \to \mathbf{P}\mathcal{D},$$

and the closure operator of the information system as a function

$$\theta : \mathbf{P}\mathcal{D} \rightarrow \mathbf{P}\mathcal{D},$$

such that $S(t) = S_t = \theta(S_t)$ for each instant t.

Now S and θ are functions between ordered sets, and so may be monotonic or not. We say that monotonicity of S with increasing time is *temporal* monotonicity of the agent's attitudes; that is, the agent's mental states exhibit temporal monotonicity if they are *cumulative*, if $S_t \subseteq S_{t'}$ whenever $t \leq t'$. *Logical* monotonicity is the usual property of deductive closure functions θ; the set of conclusions grows monotonically with increasing sets of axioms, that is, $\theta(X) \subseteq \theta(X')$ whenever $X \subseteq X'$. Thus temporal or logical nonmonotonicity occurs when the agent's characterization employs nonmonotonic functions for S or instead of θ.

The idea that reasoning may be nonmonotonic is very old, for in almost all familiar situations the attitudes of agents change nonmonotonically over time; that is, the function S is temporally nonmonotonic in ordinary situations.

Rational agents provide a good example of logical nonmonotonicity. When meanings are constant and environments ignored, we may assume the existence of a transition function

$$\tau : \mathcal{I} \rightarrow \mathcal{I}$$

such that

$$\tau(S_t) = S_{t+1}.$$

In rational agents, such functions, considered as functions $\tau : \mathbf{P}\mathcal{D} \rightarrow \mathbf{P}\mathcal{D}$ over sets of attitudes, are ordinarily nonmonotonic, as they typically involve choosing actions on the basis of the transitive closures of incomplete sets of likelihoods and preferences, so that larger states, which may contain more preferences and likelihoods, may yield different maxima. Thus τ is logically nonmonotonic. Of course, τ also describes the agent's (possibly nonmonotonic) temporal evolution, and this encourages a degree of confusion between the two sorts of nonmonotonicity.

References

Abelson, H., Sussman, G. J., and Sussman, J., 1985. *Structure and Interpretation of Computer Programs*, Cambridge: MIT Press.

Arrow, K. J., 1963. *Social Choice and Individual Values*, second edition, New Haven: Yale University Press.

Arrow, K. J., 1974. *The Limits of Organization*, New York: Norton.

Arrow, K. J., and Hahn, F. H., 1971. *General Competitive Analysis*, Amsterdam: North-Holland.

Baron, J., 1985. *Rationality and Intelligence*, Cambridge: Cambridge University Press.

Barth, J., 1967. *The Sot-Weed Factor* (rev. ed.), New York: Doubleday, 238-239.

Barwise, J., 1985. Model-theoretic logics: background and aims, *Model-Theoretic Logics* (J. Barwise and S. Feferman, eds.), New York: Springer-Verlag, 3-23.

Barwise, J., and Perry, J., 1983. *Situations and Attitudes*, Cambridge: MIT Press.

Batali, J., 1985. A computational theory of rational action (draft), Cambridge: MIT AI Lab.

Becker, G. S., 1976. *The Economic Approach to Human Behavior*, Chicago: University of Chicago Press.

Berger, J. O., 1985. *Statistical Decision Theory and Bayesian Analysis (second edition)*, New York: Springer-Verlag.

Berger, P. L., and Luckmann, T., 1966. *The Social Construction of Reality: A treatise in the sociology of knowledge*, New York: Doubleday.

Black, D., 1963. *The Theory of Committees and Elections*, Cambridge: Cambridge University Press.

Brunsson, N., 1985. *The Irrational Organization: Irrationality as a Basis for Organizational Action and Change*, Chichester: Wiley.

Buchanan, J. M., and Tullock, G., 1962. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*, Ann Arbor: University of Michigan Press.

Carbonell, J. G., 1986. Derivational analogy: a theory of reconstructive problem solving and expertise acquisition, *Machine Learning: An Artificial Intelligence Approach, Volume 2* (R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, eds.), Los Altos: Morgan-Kaufmann, 371-392.

Chapman, D., 1987. Planning for conjunctive goals, *Artificial Intelligence*, V. 32, 333-377.

Chapman, D., and Agre, P. E., 1987. Abstract reasoning as emergent from concrete activity, *Reasoning about Actions and Plans* (M. P. Georgeff and A. L. Lansky, eds.), Los Altos, CA: Morgan Kaufmann, 411-424.

Charniak, E., and McDermott, D., 1985. *Introduction to Artificial Intelligence*, Reading: Addison-Wesley.

Cherniak, C., 1986. *Minimal Rationality*, Cambridge: MIT Press.

Cook, S. A., 1983. An overview of computational complexity, *C.A.C.M. 26*, 401-408.

Davis, L. H., 1979. *Theory of Action*, Englewood Cliffs: Prentice-Hall.

Davis, M., 1981. Obvious logical inferences, *Seventh IJCAI*, 530-531.

Davis, R., 1980. Meta-rules: reasoning about control, MIT AI Laboratory, Memo 576.

Debreu, G., 1959. *Theory of Value: an axiomatic analysis of economic equilibrium*, New York: Wiley.

de Kleer, J., 1986. An assumption-based TMS, *Artificial Intelligence* **28**, 127-162.

de Kleer, J., Doyle, J., Steele, G. L. Jr., and Sussman, G. J., 1977. AMORD: Explicit control of reasoning, *Proc. ACM Conference on AI and Programming Languages*, Rochester, New York.

Dennett, D. C., 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*, Cambridge: MIT Press.

de Sousa, R., 1971. How to give a piece of your mind: or, the logic of belief and assent, *Rev. of Metaphysics* **XXV**, 52-79.

de Sousa, R., 1987. *The Rationality of Emotion*, Cambridge: MIT Press.

Doyle, J., 1979. A truth maintenance system, *Artificial Intelligence* **12**, 231-272.

Doyle, J., 1980. A model for deliberation, action, and introspection, Cambridge: MIT Artificial Intelligence Laboratory, TR-581.

Doyle, J., 1982. The foundations of psychology, Pittsburgh: Department of Computer Science, Carnegie-Mellon University, report 82-149.

Doyle, J., 1983a. Some theories of reasoned assumptions: an essay in rational psychology, Pittsburgh: Carnegie-Mellon University, Department of Computer Science, report 83-125.

Doyle, J., 1983b. What is rational psychology? toward a modern mental philosophy, *AI Magazine*, V. 4, No. 3, 50-53.

Doyle, J., 1983c. The ins and outs of reason maintenance, *Eighth International Joint Conference on Artificial Intelligence*, 349-351.

Doyle, J., 1983d. A society of mind: multiple perspectives, reasoned assumptions, and virtual copies, *Eighth International Joint Conference on Artificial Intelligence* 309-314.

Doyle, J., 1983e. Admissible state semantics for representational systems, *IEEE Computer*, V. 16, No. 10, 119-123.

Doyle, J., 1985. Reasoned assumptions and Pareto optimality, *Ninth International Joint Conference on Artificial Intelligence* 87-90.

Doyle, J., 1988a. Knowledge, representation, and rational self-government, *Proc. Second Conf. on Theoretical Aspects of Reasoning about Knowledge* (M. Y. Vardi, ed.), Los Altos: Morgan Kaufmann.

Doyle, J., 1988b. Big problems for artificial intelligence, *AI Magazine*, (to appear).

Doyle, J., 1988c. On universal theories of defaults, Pittsburgh: Carnegie Mellon University, Computer Science Department, TR CMU-CS-88-111.

Doyle, J., 1988d. On rationality and learning, Pittsburgh: Carnegie Mellon University, Computer Science Department, TR CMU-CS-88-122.

Doyle, J., 1988e. Similarity, conservatism, and rationality, Pittsburgh: Carnegie Mellon University, Computer Science Department, TR CMU-CS-88-123.

Dreyfus, H. L., and Dreyfus, S. E., 1985. *Mind Over Machine*, New York: Macmillan.

Drucker, P. F., 1985. *Innovation and Entrepreneurship: Practice and Principles*, New York: Harper and Row.

Ellis, A. and Harper, R. A., 1961. *A Guide to Rational Living*, Englewood Cliffs: Prentice-Hall.

Ellis, B., 1979. *Rational Belief Systems*, Totowa: Rowman and Littlefield.

Elster, J., 1979. *Ulysses and the Sirens: Studies in rationality and irrationality*, Cambridge: Cambridge University Press.

Elster, J., 1983. *Sour Grapes: Studies in the Subversion of Rationality*, Cambridge: Cambridge University Press.

Fagin, R., Ullman, J. D., and Vardi, M. Y., 1983. On the semantics of updates in databases, *Proc. Second ACM SIGACT-SIGMOD*, 352-365.

Fahlman, S. E., 1979. *NETL: A System for Representing and Using Real World Knowledge*, Cambridge: MIT Press.

Festinger, L., 1957. *A Theory of Cognitive Dissonance*, Stanford: Stanford University Press.

Festinger, L., 1983. *The Human Legacy*, New York: Columbia University Press.

Fikes, R. E., and Nilsson, N. J., 1971. STRIPS: a new approach to the application of theorem proving to problem solving, *Artificial Intelligence* **2**, 189-208.

Fodor, J. A., 1983. *The Modularity of Mind: An Essay on Faculty Psychology*, Cambridge: MIT Press.

Frankfurt, H., 1971. Freedom of the will and the concept of a person, *Journal of Philosophy* **68**, 5-20.

Gärdenfors, P., 1988. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*, (to appear).

Garey, M. R., and Johnson, D. S., 1979. *Computers and Intractability: a guide to the theory of NP-completeness*, San Francisco: W. H. Freeman.

Gazzaniga, M. S., 1985. *The Social Brain: Discovering the Networks of the Mind*, New York: Basic Books.

Geertz, C., 1983. Common sense as a cultural system, *Local Knowledge*, New York: Basic Books, 73-93.

Genesereth, M. R., and Nilsson, N. J., 1987. *Logical Foundations of Artificial Intelligence*, Los Altos: Morgan Kaufmann.

Goldman, A. I., 1970. *A Theory of Human Action*, Princeton: Princeton University Press.

Goodwin, J. W., 1987. A theory and system for non-monotonic reasoning, Ph.D. thesis, Department of Computer and Information Science, Linköping University. Linköping Studies in Science and Technology, No. 165.

Grabiner, J. V., 1986. Computers and the nature of man: a historian's perspective on controversies about artificial intelligence, *Bulletin of the American Mathematical Society* **15**, 113-126.

Haack, S., 1978. *Philosophy of Logics*, Cambridge: Cambridge University Press.

Hamming, R. W., 1962. *Numerical Methods for Scientists and Engineers*, New York: McGraw-Hill, Chapter $N + 1$, 394-401.

Hanks, S., and McDermott, D., 1985. Temporal reasoning and default logics, New Haven: Department of Computer Science, Yale University, report 430.

Harel, D., 1984. Dynamic logic, *Handbook of Philosophical Logic* (D. Gabbay and F. Guenthner, eds.), Dordrecht: Reidel, Vol. II, 497-604.

Harel, D., 1987. *Algorithmics: The Spirit of Computing*, Reading: Addison-Wesley.

Harman, G., 1973. *Thought*, Princeton: Princeton University Press.

Harman, G., 1986. *Change of View: Principles of Reasoning*, Cambridge: MIT Press.

Harper, W. L., 1976. Rational belief change, Popper functions, and counterfactuals, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Vol. 1, (W. L. Harper and C. A. Hooker, eds.), Dordrecht: Reidel, 73-115.

Hayes, P. J., 1973. The frame problem and related problems in artificial intelligence, *Artificial and Human Thinking* (A. Elithorn and D. Jones, eds.), San Francisco: Josey-Bass, 45-49. Reprinted in *Readings in Artificial Intelligence* (B. L. Webber and N. J. Nilsson, eds.), Los Altos: Morgan-Kaufmann (1981), 223-230.

Hayes, P. J., 1974. Some problems and non-problems in representation theory, *Proc. Conf. Artificial Intelligence and Simulation of Behavior*, 63-79. Reprinted in *Readings in Knowledge Representation* (R. J. Brachman and H. J. Levesque, eds.), Los Altos: Morgan-Kaufmann (1985), 4-22.

Hirschman, A. O., 1982. *Shifting Involvements: Private interest and public action*, Princeton: Princeton University Press.

Howard, R. A., 1980. An assessment of decision analysis, *Operations Research*, V. 28, No. 1, 4-27.

James, W., 1897. *The Will to Believe and other essays in popular philosophy*, New York: Longmans, Green and Co.

Jaynes, J., 1976. *The Origin of Consciousness in the Breakdown of the Bicameral Mind*, Boston: Houghton-Mifflin.

Jeffrey, R. C., 1983. *The Logic of Decision*, second edition, Chicago: University of Chicago Press.

Karp, R. M., 1986. Combinatorics, complexity, and randomness, *C.A.C.M. 29*, 98-109.

Knuth, D. E., and Moore, R. N., 1975. An analysis of alpha-beta pruning, *Artificial Intelligence* **6**, 293-326.

Kyburg, H. E., Jr., 1970. *Probability and Inductive Logic*, New York: Macmillan.

Kydland, F. E., and Prescott, E. C., 1977. Rules rather than discretion: the inconsistency of optimal plans, *J. Political Economy*, Vol. 85, No. 3, 473-491.

Laird, J. E., Newell, A., and Rosenbloom, P. S., 1987. SOAR: an architeture for general intelligence, *Artificial Intelligence*, V. 33., 1-64.

Lakoff, G., 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, Chicago: University of Chicago Press.

Langlotz, C. P., Shortliffe, E. H., and Fagan, L. M., 1986. Using decision theory to justify heuristics, *Proc. Fifth National Conference on Artificial Intelligence*, 215-219.

Leibenstein, H., 1980. *Beyond Economic Man: A new foundation for microeconomics*, second ed., Cambridge: Harvard University Press.

Lenat, D., Davis, R., Doyle, J., Genesereth, M., Goldstein, I., and Shrobe, H., 1983. Reasoning about reasoning, *Building Expert Systems* (D. Waterman, R. Hayes-Roth, and D. Lenat, eds.), Reading: Addison-Wesley, 219-239.

Levesque, H. J., 1984. A logic of implicit and explicit belief, *AAAI-84*, 198-202.

Levi, I., 1967. *Gambling with Truth: an essay on induction and the aims of science*, New York: Knopf.

Levi, I., 1980. *The Enterprise of Knowledge: an essay on knowledge, credal probability, and chance*, Cambridge: MIT Press.

Levi, I., 1986. *Hard Choices: Decision making under unresolved conflict*, Cambridge: Cambridge University Press.

Lewis, D., 1973. *Counterfactuals*, Cambridge: Harvard University Press.

Lindblom, C. E., 1977. *Politics and Markets: The World's Political-Economic Systems*, New York: Basic Books.

Luce, R. D., and Raiffa, H., 1957. *Games and Decisions*, New York: Wiley.

McAllester, D. A., 1980. An outlook on truth maintenance, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, AI Memo 551.

McCarthy, J., 1958. Programs with common sense, *Readings in Knowledge Representation* (R. J. Brachman and H. J. Levesque, eds.), Los Altos: Morgan Kaufmann (1985), 300-307.

McCarthy, J., 1977. Epistemological problems of artificial intelligence, *IJCAI-77*, 1038-1044. Reprinted in *Readings in Artificial Intelligence* (B. L. Webber and N. J. Nilsson, eds.), Los Altos: Morgan-Kaufmann (1981), 459-465.

McCarthy, J., 1980. Circumscription—a form of non-monotonic reasoning, *Artificial Intelligence* **13**, 27-39. Reprinted in *Readings in Artificial Intelligence* (B. L. Webber and N. J. Nilsson, eds.), Los Altos: Morgan-Kaufmann (1981), 466-472.

McCarthy, J., 1986. Applications of circumscription to formalizing common-sense knowledge, *Artificial Intelligence* **28**, 89-116.

McCarthy, J., and Hayes, P. J., 1969. Some philosophical problems from the standpoint of artificial intelligence, *Machine Intelligence 4* (B. Meltzer and D. Michie, eds.), New York: American Elsevier, 463-502. Reprinted in *Readings in Artificial Intelligence* (B. L. Webber and N. J. Nilsson, eds.), Los Altos: Morgan-Kaufmann (1981), 431-450.

McDermott, D., 1978. Planning and acting, *Cognitive Science* **2**, 71-109.

McDermott, D., 1987. Critique of pure reason, *Computational Intelligence*, Vol. 3, No. 3, 151-160.

McDermott, D., and Doyle, J., 1980. Non-monotonic logic—I, *Artificial Intelligence* **13**, 41-72.

Maital, S., 1982. *Minds, Markets, and Money: Psychological foundations of economic behavior*, New York: Basic Books.

March, J. G., and Simon, H. A., 1958. *Organizations*, New York: Wiley.

Miller, G. A., 1986. Dismembering cognition, *One Hundred Years of Psychological Research in America* (S. H. Hulse and B. F. Green, Jr., eds.), Baltimore: Johns Hopkins University Press, 277-298.

Minsky, M., 1961. Steps towards artificial intelligence, *Computers and Thought* (E. A. Feigenbaum and J. Feldman, eds.), New York: McGraw-Hill, 406-450.

Minsky, M., 1965. Matter, mind, and models, *Proc. of the IFIP Congress*, 45-49. Reprinted in M. Minsky (ed.), *Semantic Information Processing*, MIT Press, Cambridge, (1965), 425-432.

Minsky, M., 1975. A framework for representing knowledge, *The Psychology of Computer Vision* (P. Winston, ed.), New York: McGraw-Hill. Appendix in MIT AI Laboratory Memo 306.

Minsky, M., 1986. *The Society of Mind*, New York: Simon and Schuster.

Moore, R. C., 1983. Semantical considerations on nonmonotonic logic, *Eighth International Joint Conference on Artificial Intelligence*, 272-279.

Mueller, D. C., 1979. *Public Choice*, Cambridge: Cambridge University Press.

Mundell, R. A., 1968. *Man* and *Economics*, New York: McGraw-Hill. See especially chapter 18.

Nagel, T., 1979. The absurd, *Mortal Questions*, Cambridge: Cambridge University Press, 11-23.

Newell, A., and Simon, H. A., 1963. GPS, a program that simulates human thought, *Computers and Thought* (E. A. Feigenbaum and J. Feldman, eds.), New York: McGraw-Hill, 279-293.

Nowakowska, M., 1973. *Language of Motivation and Language of Actions*, The Hague: Mouton.

Ostwald, M., 1962. Aristotle, *Nichomachian Ethics* (translator), Indianapolis: Bobbs-Merrill.

Pareto, V., 1927. *Manual of Political Economy* (tr. A. S. Schwier, ed. A. S. Schwier and A. N. Page), New York: Kelley, 1971.

Pascal, B., 1662. *Pensées sur la religion et sur quelques autres sujets* (tr. M Turnell), London: Harvill, 1962.

Pearl, J., 1984. *Heuristics: Intelligent search strategies for computer problem solving*, Reading: Addison-Wesley.

Pears, D., 1984. *Motivated Irrationality*, Oxford: Oxford University Press.

Peck, M. S., 1978. *The Road Less Travelled: A new psychology of love, traditional values and spiritual growth*, New York: Simon and Schuster.

Polya, G., 1962. *Mathematical Discovery: On understanding, learning, and teaching problem solving*, Volume 1, New York: Wiley.

Polya, G., 1965. *Mathematical Discovery: On understanding, learning, and teaching problem solving*, Volume 2, New York: Wiley.

Quine, W. V., 1953. Two dogmas of empiricism, *From a Logical Point of View*, Cambridge: Harvard University Press.

Quine, W. V., 1970. *Philosophy of Logic*, Englewood Cliffs: Prentice-Hall.

Quine, W. V., and Ullian, J. S., 1978. *The Web of Belief*, second edition, New York: Random House.

Raiffa, H., 1968. *Decision Analysis: Introductory Lectures on Choices Under Uncertainty*, Reading, MA: Addison-Wesley.

Rawls, J., 1971. *A Theory of Justice*, Cambridge: Harvard University Press.

Reinfrank, M., 1985. An introduction to non-monotonic reasoning, Universität Kaiserlautern, Informatik, SEKI Memo 85-02.

Reiter, R., 1980. A logic for default reasoning, *Artifical Intelligence* **13**, 81-132.

Rescher, N., 1964. *Hypothetical Reasoning*, Amsterdam: North Holland.

Rorty, R., 1979. *Philosophy and the Mirror of Nature*, Princeton: Princeton University Press.

Rosenberg, N., and Birdzell, L. E. Jr., 1986. *How the West Grew Rich: The economic transformation of the industrial world*, New York: Basic Books.

Rosenschein, S. J., and Kaelbling, L. P., 1986. The synthesis of digital machines with provable epistemic properties, *Theoretical Aspects of Reasoning about Knowledge* (J. Y. Halpern, ed.), Los Altos: Morgan Kaufmann, 83-98.

Russell, B., 1930. *The Conquest of Happiness*, New York: Liveright.

Sacerdoti, E. D., 1974. Planning in a hierarchy of abstraction spaces, *Artificial Intelligence 5*, 115-135.

Sacerdoti, E. D., 1977. *A Structure for Plans and Behavior*, New York: American Elsevier.

Savage, L. J., 1972. *The Foundations of Statistics*, 2nd rev. ed., New York: Dover.

Schelling, T. C., 1984a. The intimate contest for self-command, *Choice and Consequence: Perspectives of an errant economist*, Cambridge: Harvard University Press, 57-82.

Schelling, T. C., 1984b. The mind as a consuming organ, *Choice and Consequence: Perspectives of an errant economist*, Cambridge: Harvard University Press, 328-346.

Schön, D. A., 1983. *The Reflective Practitioner: How Professionals Think in Action*, New York: Basic Books.

Schumpeter, J. A., 1934. *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle* (R. Opie, tr.), Cambridge: Harvard University Press.

Scitovsky, T., 1976. *The Joyless Economy: An inquiry into human satisfaction and consumer dissatisfaction*, Oxford: Oxford University Press.

Scott, D. S., 1982. Domains for denotational semantics, *International Conference on Automata, Languages, and Programming*.

Searle, J. R., 1983. *Intentionality: An Essay in the Philosophy of Mind*, Cambridge: Cambridge University Press.

Seidenfeld, T., Kadane, J., and Schervish, M., 1986. Decision theory without ordering, in preparation.

Shafer, G. R., 1976. *A Mathematical Theory of Evidence*, Princeton: Princeton University Press.

Shoham, Y., 1988. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence*, Cambridge: MIT Press.

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

Simon, H. A., 1982. *Models of Bounded Rationality, Volume 2: Behavioral Economics and Business Organization*, Cambridge: MIT Press.

Smith, B. C., 1982. Reflection and semantics in a procedural language, Cambridge: Laboratory for Computer Science, Massachusetts Institute of Technology, TR-272.

Smith, D. E., 1985. Controlling inference, Stanford: Department of Computer Science, Stanford University, Ph.D. thesis.

Stalnaker, R. C., 1984. *Inquiry*, Cambridge: MIT Press.

Stefik, M. J., 1980. Planning with constraints, Stanford University, Computer Science Department, Report STAN-CS-80-784.

Sternberg, R. J., 1986. Intelligence is mental self-government, *What is Intelligence? Contemporary Viewpoints on its Nature and Definition* (R. J. Sternberg and D. K. Detterman, eds.), Norwood, NJ: Ablex, 141-148.

Stigler, G. J., and Becker, G. S., 1977. De gustibus non est disputandum, *American Economic Review*, Vol. 67, No. 2, 76-90.

Sussman, G. J., 1975. *A Computer Model of Skill Acquisition*, New York: American Elsevier.

Sussman, G. J., and G. L. Steele Jr., 1980. CONSTRAINTS—A language for expressing almost-hierarchical descriptions, *Artificial Intelligence* **14**, 1-39.

Thaler, R. H., and Shefrin, H. M., 1981. An economic theory of self-control, *J. Political Economy*, Vol. 89, No. 2, 392-406.

Thomason, R. H., 1979. Some limitations to the psychological orientation in semantic theory, mimeo, Pittsburgh: University of Pittsburgh.

Thomason, R. H., 1987. The context-sensitivity of belief and desire, *Reasoning about Actions and Plans* (M. P. Georgeff and A. L. Lansky, eds.), Los Altos: Morgan Kaufmann, 341-360.

Thurow, L. C., 1983. *Dangerous Currents: The state of economics*, New York: Random House.

Tinbergen, N., 1951. *The Study of Instinct*, Oxford: Clarendon Press.

Touretzky, D. S., 1986. *The Mathematics of Inheritance Systems*, London: Pitman.

Touretzky, D., Horty, J., and Thomason, R., 1987. A clash of intuitions: the current state of nonmonotonic multiple inheritance systems, *Ninth International Joint Conference on Artificial Intelligence*, 476-482.

Truesdell, C., 1984. Is there a philosophy of science? *An Idiot's Fugitive Essays on Science: Methods, Criticism, Training, Circumstances*, New York: Springer-Verlag, 471-502.

Ullman, S., 1983. Visual routines, MIT AI Lab, AI Memo 723.

Valiant, L. G., 1984. A theory of the learnable, *Comm. A.C.M.*, Vol. 18, No. 11, 1134-1142.

von Neumann, J., and Morgenstern, O., 1944. *Theory of Games and Economic Behavior*, Princeton: Princeton University Press.

Waldinger, R., 1977. Achieving several goals simultaneously, *Machine Intelligence 8* (E. W. Elcock and D Michie, eds.), Chichester: Ellis Horwood, 94-136. Reprinted in *Readings in Artificial Intelligence* (B. L. Webber and N. J. Nilsson, eds.), Los Altos: Morgan-Kaufmann (1981), 250-271.

Weyhrauch, R. W., 1980. Prolegomena to a theory of mechanized formal reasoning, *Artificial Intelligence* **13**, 133-170.

Williamson, O. E., 1975. *Markets and Hierarchies: Analysis and Antitrust Implications: A study of the economics of internal organization*, New York: Free Press.

Winograd, T., and Flores, F., 1986. *Understanding Computers and Cognition: A New Foundation for Design*, Norwood, NJ: Ablex.

Yates, B. T., 1985. *Self-Management: The Science and Art of Helping Yourself*, Belmont, California: Wadsworth.

Glossary of Some Key Terms

These glosses are meant to suggest the senses in which we use these terms, but should not be taken as strict definitions, since many are primitive terms whose exact character stems from the roles they play in the formal treatment.

Accommodation: A change of legal state undertaken in order to incorporate some attitudes.

Attitudes: Mental stances towards things, such as hopes that certain conditions are true, or hatreds of certain activities. While humans exhibit many sorts of attitudes, we are mainly concerned with beliefs, desires, and intentions.

Belief: The attitude toward a proposition of believing it to be true, believing the proposition describes the actual world. Relative beliefs are called likelihoods.

Conservative: Changes of state that are as small as possible.

Constitution: The organization of the agent's legal states and actions, as expressed by the underlying constitutive logic (or psycho-logic) and constitutive attitudes.

Desire: The attitude toward a proposition of wanting it to be true. Sometimes also called wants or goals. Relative desires are called preferences.

External: A part of the world viewed as part of the agent's environment.

Habitual: Actions or steps of reasoning taken by the agent without thinking about whether to do them.

Inference: See reasoning.

Intent: The attitude toward a proposition of intending to act to make the proposition true. Sometimes also called plans or goals or aims. Relative intentions are called priorities.

Internal: A part of the world viewed as part of the agent.

Legal states: States closed and consistent according to the agent's constitutive logic and satisfying all the constitutive attitudes they contain.

Likelihood: Comparative belief that one proposition is likelier to be true than another.

Logic: A formal system describing the internal structure of some subject. In logical reasoning, the set of attitudes adopted and abandoned reflect those mentioned in inference rules of the logic.

Preference: Comparative desire for one proposition to be true more than another.

Priority: Comparative intention to make one proposition true before another.

Proposition: A set of posible worlds. For example, the set of exactly those worlds in which some condition is true. A proposition holds, or is true, if it contains the actual world.

Rational: Ideally, an action is rational if the action is of maximal expected utility among all the actions available to an agent at some instant.

Reasonable: Actions humans find acceptable or normal. Reasonability may differ from rationality, especially when the agent's attitudes conflict.

Reasoning: The process of changing one's attitudes. This may involve adopting or abandoning attitudes of any type.