Reset reproduction of CMU Computer Science report CMU-CS-83-125 (only fonts and pagination differ). Versions of 1982 and 1983 differ only slightly. Revised and abbreviated version published in *Fundamenta Informaticae*, V. 20, Nos. 1–3 (1994), pp. 35–73. Reprinted April 2006. Current address: NCSU Department of Computer Science, Raleigh, North Carolina.

Some Theories of Reasoned Assumptions

An essay in rational psychology

Jon Doyle

Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

December 12, 1982
Revised May 12, 1983

This research was supported by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionics Laboratory under Contract F33615-81-K-1539. 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.

Abstract

We examine several formulations of the common practice of jumping to conclusions when actions demand decisions but solid knowledge fails. This practice permeates artificial intelligence, where systems assume many conclusions automatically as defaults simply because the questions they decide are known to occur frequently, and where other assumptions are formulated and adopted only when ignorance stalls action. After developing the motivations and general nature of these inferences, we introduce a formal basis for describing them. This formulation allows separate introduction of the several ideas involved, and so facilitates characterization of some important combinations and some previous proposals. Initial results are proved about these theories, including the aptness of the formal notions with respect to the intuitive motivations. Benefits of this formulation include an indication of the ways notions from logic and metamathematics can enter into psychologies without subscribing to all of logic or metamathematics, an indication of the importance of conservation of mental states in the description of psychologies, and formal and intuitive relations between the approach of reasoned assumptions and its popular alternatives, deductivism and Bayesianism.

Acknowledgments

My debts, obvious and hidden, are many. As always, JOSEPH SCHATZ helped more than I can say. JON BARWISE, JOHANNES BRAHMS, JAIME CARBONELL, JOHAN DE KLEER, MERRICK FURST, CLARK GLYMOUR, JUSSI KETONEN, ROBERT LADDAGA, DAVID MCALLESTER, JOHN MCCARTHY, DREW MCDERMOTT, MARVIN MINSKY, ROBERT MOORE, ALLEN NEWELL, JOANN ORDILLE, RAYMOND REITER, DANA SCOTT, ROBERT STALNAKER, GERALD SUSSMAN, PETER SZOLOVITS, and RICHMOND THOMASON all helped in different ways. JAN ZUBKOFF graciously did the initial typing. The Computer Science department of Carnegie-Mellon University greatly aided my writing.

Note to the Reader

To keep this paper to a reasonable length, I assume some mathematical sophistication and acquaintance. Passage of time will, I hope, permit a more relaxed and thorough development. The symbols used are collected and glossed at the end of the paper. The headings DEFINITION, THEOREM, COROLLARY, LEMMA, and CONJECTURE serve their usual purposes, and are numbered in a single sequence. Some proofs have been omitted. The heading QUESTION often indicates an issue calling for proper formulation rather than a simple answer.

Analysis of Contents

Elemental Theories

Logical Theories

Attitudinal Theories

Evolutionary Theories

Probabilistic Theories

IV. Related Theories of Reasoned Assumptions

V. Conclusion

Meditation

Ging heut morgens übers Feld,
Tau noch auf den Gräsern hing;
Sprach zu mir der lustge Fink:
„Ei, du! Gelt? Guten Morgen! Ei gelt? Du!
Wird’s nicht eine schöne Welt? schöne Welt!?
Zink! Zink! schön und flink!
Wie mir doch die Welt gefällt!“

Auch die Glockenblum am Feld
Hat mir lustig, guter Ding
Mit dem Glöckchen kling, kling,
Ihren Morgengruss geschellt:
„Wird’s nicht eine schöne Welt? schöne Welt!?
Kling! Kling! Schönes Ding!
Wie mir doch die Welt gefällt! Hei—a!“

Und da fing im Sonnenschein
Gleich die Welt zu funkeln an;
Alles, alles, Ton und Farbe gewann in Sonnenschein!
Blum und Vogel, gross und klein!
Guten Tag, guten Tag! Ist’s nicht eine schöne Welt?
Ei du! Gelt? Schöne Welt!?

Nun fängt auch mein Glück wohl an?!
Nein! Nein! Das ich mein, mir nimmer blühen kann!

G. Mahler, *Lieder eines fahrenden Gesellen*

L’adoration de la terre;
Le sacrifice

I. Stravinsky, *Le Sacre du Printemps*

Ich will nur dir zu Ehren leben,
Mein Heiland, gib mir Kraft und Mut,
Dass es mein Herz recht eifrig tut.
Stärke mich, deine Gnade würdiglich und mit Danken zu erheben.

J. S. Bach, *Weihnachts-Oratorium*

for Norma Charlotte Schleif Miller

and in memory of

Henry Aloysius Miller,
Meta Augusta Wilhelmina Frederica Enters Doyle,
and
Lawrence Henry Doyle

I. Introduction

§**1.** Recently, increasing attention has been directed toward problems of providing mathematical formulations and semantics for some of the inferential systems developed in informal, practical terms within artificial intelligence. The mathematical approach has seldom been popular, for except in rare cases like the treatise of MINSKY and PAPERT on perceptrons,[1] mathematical formulations have lacked force, have seemed mere importation of formalism without true understanding of the important problems to be addressed. Artificial intelligence has long been a field for formulation, where ubiquitous problems have defied stable statements.[2] In this setting it is natural that outsiders experience difficulties in looking for important problems, since the set changes with every observation! Fortunately, circumstances change, and the field has begun to develop its own mathematical formulations of some of these recognized but poorly articulated problems. These formulations may not yet be as comprehensive or compelling as those underlying the exact sciences, but they easily support optimism for more satisfying replacements in the near future.

§**2.** Our subject here is a reexamination of several proposals concerning the mathematical formulation of certain non-deductive inference patterns common in artificial intelligence. Artificial intelligence systems draw many logically peculiar conclusions, and the question has been raised of whether there are coherent patterns among and justifications for these conclusions. Traditionally the nature of these non-deductive inferences has been hidden behind the slogan "heuristic," but realization of the widespread use of particular patterns has prompted their formulation and explication as important problems for the field. With the works of MCDERMOTT, REITER, MCCARTHY and others advances have been made, but deficiencies in understanding persist. Motivationally, these efforts articulate some of the intuitions underlying the field's practice, but apparently not well enough to communicate these intuitions and their significance to some inside and many outside the field. Formally, these proposals exhibit some similiarities so that they all seem to approximate a single answer to a single problem, but they differ widely in detail, and no exact characterization of their differences or individual powers has been given. My purpose here is to pursue of the common project of these proposals by summarizing and extending the motivations underlying the subject inference patterns. I present a common mathematical basis in which I analyze, compare, and extend the previous proposals. Hopefully this effort will benefit several audiences: artificial intelligence theoreticians and practitioners interested in better understanding their systems of study, mathematicians and philosophers interested in the formalization and motivations of these novel inferential systems, and theoretical computer scientists interested in computational questions arising in artificial intelligence.

Briefly, the problem is to formulate the common practice of jumping to conclusions when actions demand decisions but solid knowledge fails. This practice permeates artificial intelligence systems, where some assumptions are formulated and adopted only when ignorance stalls action, and where other conclusions are automatically assumed as defaults simply because the questions they decide are known to occur frequently.

[1] [MINSKY AND PAPERT 1969]
[2] Compare [MINSKY 1962].

II. A Problem, an Approach, and a Solution

§**3.** Life calls for action, and to act we must decide what to do in ignorance of our true circumstances, capabilities, and their consequences. We may decide using information about the consequences of our actions in our circumstances, but are our imagined circumstances our actual ones? Will the actions we perform be the ones we attempt? Will unforeseen interferences prevent the expected consequences?

The skeptical challenge to the possibility of accurate knowledge of the world has never taken lives; neither does it halt work in artificial intelligence. One does the best one can. To do so, however, one must be able to admit and correct one's errors, and this consideration influences the design of artificial agents as well as the conduct of life. But in artificial intelligence, incompleteness of information looms larger than inaccuracy. Certainly to admit error requires one to judge oneself ignorant in the past, but in principle, as in politics, it still allows one to maintain an opinion on every subject and to assert correctness of all one's *current* opinions. Nevertheless, there are several severe obstacles to designing agents whose information about the world is complete, even if inaccurate. The first obstacle is that apparently accurate complete axiomatizations of the world may not exist. As in arithmetic, we may be able to state our basic premises, prove their incompleteness, and convince ourselves that any completion we attempt will involve inconsistencies. Since we expect error anyway, this need not be a strong deterrent, but it does suggest that completion for its own sake is unwarranted. The second obstacle, one far more serious, is the feasibility of using complete information even if we attain it. Present-day computation is recursive computation, and an undecidable complete axiomatization is no better than an incomplete axiomatization. Worse still, even decidable theories may be intractable, for short theorems may have very long shortest proofs, and answering single questions may involve exploring very many potential proofs. Since time will not stand still while the agent attempts to compute the answers it needs, theoretical completeness of the agent's information can be a practical fraud.[3] The third obstacle is simple unavailability of complete axiomatizations of the world. Even looking to the sciences, one finds searches for new laws to fill old gaps, and artificial intelligence has spawned an entire discipline devoted simply to elicting all the facts about the world employed by experts but unseen in the literature. It seems methodologically unwise to postpone work on artificial intelligence until science completes its inquiry, unless one does not want artificial agents at all.

§**4.** If we accept the challenge of acting without complete information, we must adopt some approach which allows both cognizance of incompletenesses and means for overcoming them. To act without awareness of one's clear limitations is blind stupidity, yet ignorance need not paralyze either. One approach out of several possible approaches has dominated work in artificial intelligence, and this is the practice of jumping to conclusions. Many of the beliefs assumed by default or for heuristic value in mechanized agents are cases of what WILLIAM JAMES called the "will to believe."[4] JAMES was, like PASCAL before him,[5] concerned primarily with questions of momentous, eternal import rather than the questions of mundane, temporal expediency common in routine thought and action, but the idea is the same. One judges, either at the moment or in advance, that it is better to adopt a stance on some issue and risk error than to take no stance at all. The position taken need not be precedential, for along with recognition of the possibility of error, we may also recognize that other or later circumstances raising similar questions may be decided differently. In TUKEY'S phrase, we often decide to act for the time being as if something were the case, rather than simply deciding something is the case.[6] But precedential or not, the approach of adopting stances carries with it a commitment to correcting mistakes when they come to light. As JAMES puts it, we might resolve to *Believe Truth!* and to *Shun Error!*, but the preceding suggests the latter resolve is best realized as conscientious correction rather than intellectual cowardice.

The approach of jumping to conclusions finds many followers in artificial intelligence, but in other disciplines studying intelligent action other approaches attract the most attention. Of these, the most influential is

[3] The importance of computational feasibility in artificial intelligence is easily underestimated. See [MINSKY 1963], [MINSKY AND PAPERT 1969], [MINSKY 1975, APPENDIX], [RABIN 1974] and [GAREY AND JOHNSON 1979] for illuminating discussions of this issue.

[4] [JAMES 1897]

[5] [PASCAL 1662]

[6] [TUKEY 1960]

subjective Bayesian decision theory.[7] There is much to be said about this alternative, but now is not our time to do so. For the moment, our task is laying out the motivations, nature, and formalizations of jumping to conclusions as it is practiced in artificial intelligence. Later we adduce some connections of this approach with subjective Bayesian decision theory which may illuminate their respective computational conveniences and difficulties.

§**5.** Once we decide to face the problem of incomplete information by deliberately adopting stances when necessary, we must also face subsidiary problems: how can we tell when we should jump to conclusions, and which ones we should settle on? Once again, the practice of artificial intelligence supplies an approach to these problems, although possibly not the only or best one. This approach uses *ratiocinative rules of thumb* to guide the adoption of assumptions. These rules state which sorts of circumstances call for making assumptions, which assumption to try first, and when and how to revise one's opinion to other assumptions. That is, ratiocinative rules of thumb serve three functions: to enforce anti-agnosticism when appropriate, to indicate a sequence in which alternatives should be tried, and to recognize circumstances which might call into question one or more of the alternatives. These rules are often embodied as general or schematic "defaults" used in drawing conclusions, rules tolerant of exceptions in that specific conclusions are defeasible on a case-by-case basis without affecting the operation of the general rule itself.[8] For example, one might decide to believe that ordinarily, every bird can fly. This decision might be carried out by a ratiocinative rule of thumb which infers of each individual bird considered that it can fly. This rule would continue to make such assumptions about newly considered birds even after particular flightless birds are recognized and their corresponding assumptions abandoned.

§**6.** Ratiocinative rules of thumb form an approach, but in the absence of precise criteria for their application and interpretation are not themselves a solution to the problem of adopting assumptions. These rules guide inferences, but are not inference rules in the usual logical sense, since questions of soundness do not enter into the discussion. The point of the rules is, after all, to be unsound, to draw conclusions not strictly entailed by their grounds. We propose a solution based on interpreting these rules of thumb as expressions of *ratiocinative desires*, regarding actions of jumping to conclusions as actions satisfying the ratiocinative desires. In these terms we can cast the anti-agnostic, sequencing, and defeasibility functions of ratiocinative rules of thumb as components of the following "syllogism."

[7] See, for example, [SAVAGE 1972], [LEVI 1967], and [LUCE AND RAIFFA 1957].

[8] Widespread use of the term "default" follows MINSKY'S influential discussion in [MINSKY 1975]. [REITER 1978] identifies a variety of appearances of this notion in the practice of artificial intelligence.

I am in circumstances A.

When I am in circumstances A, I prefer being decided about Q to being undecided about Q.

When I am in circumstances A and am undecided about Q, I prefer adopting stand C to adopting stand B.

Therefore, I should adopt stand C.

We combine and abbreviate the two implicative components of such syllogisms in rules written as

$$A \setminus\setminus B \mid\!\!\vdash C,$$

rules read as "A without B gives C" and informally interpreted as "if A obtains, and B does not, then adopt C." These rules are called *reasons* for their conclusions, and the conclusions drawn are called *reasoned assumptions*. This terminology diverges somewhat from standard usage, in which one calls the premises of the inference the reasons for the conclusion. We always refer to the premises as premises, hypotheses, antecedents, presuppositions, etc., and reserve the term reason for the inference step that connects premises with conclusions. We do not insist on particular forms for what may enter into these rules; as well as logical statements of conditions, other sorts of mental components can be accommodated, as detailed in the following formal treatment.

We base our interpretations of these encapsulated syllogisms on maximization of the utility of mental states. Utility maximization has, with some justification, gained a bad name as a descriptive theory of human behavior, but our purpose here is supplying normative theories where none exist, rather than attempting to accommodate logic to the behavior of some species of agent. Indeed, as we see later, many of the maximizations of interest are all small, independent, local decisions, easily implemented and suffering from few of the difficulties arising in the more comprehensive maximizations of subjective Bayesian decision theory. As is familiar from classical decision theory, there may be several distinct interpretations of maximal utility. Moreover, ratiocinative rules of thumb can conflict on cases. We hold all desires to be incomparable without"higher-level" desires which recommend satisfying one ratiocinative desire before another. This means we seek ways to live with conflicts rather than insist on their resolution, and that maximixing utility involves maximizing the set of ratiocinative desires satisfied by mental states. For example, suppose we choose to employ the two general rules

$$\forall x \ [\textit{Republican}(x) \setminus\setminus \textit{Pacifist}(x) \mid\!\!\vdash \neg \textit{Pacifist}(x)]$$

and

$$\forall x \ [\textit{Quaker}(x) \setminus\setminus \neg \textit{Pacifist}(x) \mid\!\!\vdash \textit{Pacifist}(x)] \, .$$

The first of these says that Republicans should be assumed to be non-pacifists unless known to be pacifists. The second says that Quakers should be assumed to be pacifists unless known to be otherwise. When we learn that Richard Nixon is both Republican and Quaker, we can instantiate these general rules to the specific cases

$$\textit{Republican}(\textit{Nixon}) \setminus\setminus \textit{Pacifist}(\textit{Nixon}) \mid\!\!\vdash \neg \textit{Pacifist}(\textit{Nixon})$$

and

$$\textit{Quaker}(\textit{Nixon}) \setminus\setminus \neg \textit{Pacifist}(\textit{Nixon}) \mid\!\!\vdash \textit{Pacifist}(\textit{Nixon}).$$

We cannot honor both preferences at once, so if we look to satisfy as many preferences as possible, we find two alternative coherent sets of assumptions: believing Nixon a Republican Quaker pacifist, and believing him a Republican Quaker non-pacifist. Thus in accord with the interpretation of reasons as desires, we adopt a stand by picking one of these coherent sets as our set of assumptions or beliefs.

Some might find it more reasonable to suspend judgement in cases of conflicting preferences, but under our interpretation that is a poor solution to the difficulty. In adopting these reasons, we have stated our preferences, and with no further information it is needlessly irrational to forgo all satisfactions simply because a trade-off is

involved. If the decision were between consuming equally attractive donuts and bagels when there is money to purchase only one, few would counsel starvation. Nevertheless, behind the apparent questionableness of the above decisions lies an important point. Just as we may wish to state rules about when *not* to be agostic, we may also wish to state rules about when *to be* agnostic, and these may refer to otherwise conflicting preferences like those in the example. Toward this end we introduce the notion of *defeasible reasons*. Defeasible reasons allow inferences that can be defeated as a whole, rather than by simply challenging one of the particular presuppositions of the inference. One can encode defeasible reasons in reasons of the above form by introducing explicitly self-referential presuppositions (e.g. $R =$ "$A \setminus\setminus B$ or R defeated $\|\!\!\sim C''$), but it is more elegant to simply interpret *all* reasons as containing a uniform presupposition of lack of challenges. With this modification to the interpretation, we can continue to base the interpretation on utility maximization while permitting rules of agnosticism.

§**7.** This solution to the problem of adopting assumptions raises but does not address several subsidary problems. The first of these concerns the origins and justifications of those ratiocinative rules of thumb with which we choose to endow an agent or which an agent adopts of its own accord. There are several motivations for certain sorts of rules of thumb that immediately suggest themselves, and there may be more as well. The first motivation is the classical notion of statistical likelihood, where we might choose a rule of thumb because it calls for adopting the likeliest alternative. Another motivation is typicality, in which reasons specify the typical conclusion. Typicality might be the same as statistical likelihood, but there are ways of viewing it as a different notion. Another consideration in adopting a rule of thumb is the safety of error and ease of correction, that is, whether disastrous consequences follow from errors, and if not, whether the undesired consequences can be simply remedied. Finally, the weakest motivation is simple pragmatic utility, which often goes by the name of heuristic in artificial intelligence. In this case, one might adopt a ratiocinative rule of thumb simply because one is more successful by doing so.

We propose no procedure for adopting ratiocinative rules of thumb, except for those which are themselves reasoned assumptions. Instead, we separate the issues of formulating, motivating, and adopting ratiocinative rules of thumb from the issue of interpreting them once adopted. The questions of formulation and motivation of these rules of thumb arise in other terms in the study of induction, learning, and philosophy of science, and we defer discussion to that literature.[9]

§**8.** A second subsidiary problem raised by adopting assumptions is that of how to revise mistaken assumptions, of how to honor the commitment to correcting errors as they are discovered. Freshly perceived information can contradict previous beliefs, deductive inferences can bring hidden inconsistencies to light, and one sometimes decides to abandon or avoid certain attitudes because they prove embarrassing or endanger mental stability. While some humans seem happy to hide their conflicting attitudes from themselves, some consciously endure their conflicts, and some die, the usual response to this problem is to give something up, either the new information or previous attitudes, in everyday life as well as in philosophy and artificial intelligence. But prudence counsels care in abandoning one's attitudes. Their acquisition takes time and effort, and they should not be abandoned needlessly. This conservative stance has been taken repeatedly in philosophy and in artificial intelligence. QUINE christens it the "maxim of minimum mutilation," that is, when changes are necessary, one makes as small a change as possible.[10] Less articulately, artificial intelligence practice exhibits this principle in many forms, with systems of differing levels of sophistication employing varying degrees of minimality and effectiveness in revising their attitudes. We pursue this idea as the notion of *conservation* of mental state, where conservation is equivalent in all but connotation to minimal mutilation. (I prefer the term conservatism to minimal mutilation, for I would rather be called a conservative or non-conservative than a mutilator of any degree.) Thus we view acceptable revisions of mental states to be the changes that in one way or other remedy the difficulty while preserving as much of the previous state as possible. We develop this topic formally in its own right elsewhere, as it appears many techniques from the practice of artificial intelligence can be conveniently described in these terms, with many concrete measures of "amount of change" characterizing the

[9] See [MINSKY 1963], [GOODMAN 1973], [QUINE AND ULLIAN 1978], [DACEY 1978], and [LEVI 1980].
[10] [QUINE 1970]

various practical systems. One class of systems, the so-called reason maintenance systems, minimizes the set of changed reasoned assumptions (see §46), but investigation of more refined measures is just beginning.

III. Formal Theories of Reasoned Assumptions

§**9.** This chapter formalizes the intuitive notions, approaches, and solutions of the preceding. The principal problem in specifying interpretations of reasons is how they may be aggregated in spite of conflicts, and we formalize this as the problem of defining the *admissible states* of the agent holding the reasons. These admissible states constitute the "coherent" sets of interdependent reasoned assumptions sanctioned by the agent's reasons. Our method in formalization is to construct the set of admissible states from various assumptions about the constitution of the agent. Different theories of reasoned assumptions arise through the different constitutive assumptions we make (not to be confused with the reasoned assumptions made by the agent) about the agent's composition, the interpretation of individual reasons, and the aggregation of reasons and assumptions. Our first theories concern only the static side of the agent, but later we consider further constitutive assumptions about conservatism of the agent's state changes, and so arrive at evolutionary theories of reasoned assumptions.

§**10.** We first suppose that there is a domain $\mathcal{D}$ of formal or structural elements such that each possible mental state of the agent can be decomposed into elements of $\mathcal{D}$. Since we are concerned only with the structure of states vis-à-vis reasoned assumptions, we make no suppositions about the specific composition of $\mathcal{D}$. Examples of domains from artificial intelligence include the set of all sentences in some logical language, the set of all LISP data-structures (S-expressions), and the set of all "mental agents" in MINSKY and PAPERT'S society of mind.[11] As the latter example shows, we do not require that $\mathcal{D}$ has grammatical structure or is completely representational, and as the former examples show, "languages of thought" are acceptable as well. The set of admissible states $\mathscr{S}$ is thus a subset of the set of all sets of mental components: in other words, $\mathscr{S} \subseteq \mathbf{P}\,\mathcal{D}$.

§**11.** Though we need make no suppositions about the nature of the arbitrary mental component, our second constitutive assumption is that some mental components can be interpreted as reasons, as specifications for the composition of the states in which the components occur. Formally, we assume an interpretation function $\mathcal{I} : \mathcal{D} \rightarrow \mathbf{P}\,\mathbf{P}\,\mathcal{D}$ which indicates the sets of components each component sanctions, so that a set $S \subseteq \mathcal{D}$ satisfies its component specifications just in case $S \in \mathcal{I}(d)$ for each $d \in S$. We define the class of *component-admissible sets* $\mathcal{Q} \subseteq \mathbf{P}\,\mathcal{D}$ to be those sets satisfying all their components, that is,

$$\mathcal{Q} = \{S \subseteq \mathcal{D} \mid S \in \bigcap_{d \in S} \mathcal{I}(d)\}.$$

Formalizing the notion of reasons as state specifications in this way involves several simplifications. First, rather than distinguish only some state components as reasons and leave the rest uninterpreted, we interpret every state component and give each non-reason $d \in \mathcal{D}$ the trivial interpretation $\mathcal{I}(d) = \mathbf{P}\,\mathcal{D}$ that sanctions all potential states. Second, we interpret reasons as predicates of all sets of components rather than only as predicates of the admissible states, that is, to be subsets of $\mathbf{P}\,\mathcal{D}$ rather than subsets of $\mathscr{S}$. This simplification is innocuous since we can always take the interpretation of every element to exclude all inadmissible sets. Third, we take the interpretation of components to be independent of the state containing the component. While the theories examined in this paper can all be captured within this limitation, we elsewhere consider ideas from artificial intelligence better suited by state-dependent interpretations which replace $\mathcal{I} : \mathcal{D} \rightarrow \mathbf{P}\,\mathbf{P}\,\mathcal{D}$ with $\mathcal{I}' : \mathcal{D} \times \mathbf{P}\,\mathcal{D} \rightarrow \mathbf{P}\,\mathbf{P}\,\mathcal{D}$ and $\mathcal{Q}$ with $\mathcal{Q}' = \{S \subseteq \mathcal{D} \mid S \bigcap_{d \in S} \mathcal{I}'(d, S)\}$. Finally, our fourth simplification is that each state component embodies at most one reason. The extension to multiple reasons is trivial, since conjoined specifications are interpreted by intersecting their interpretations.

[11] [MINSKY 1980]

§**12.** Our third constitutive assumption is that all admissible states are component-admissible but that certain combinations of state components can never occur in admissible states even if they occur in some component-admissible states, or formally, that we can stipulate the set of admissible states $\mathscr{S}$ to be some subset of $\mathcal{Q}$, i.e. $\mathscr{S} \subseteq \mathcal{Q}$. For example, if components are sentences of a logical language, we might require every admissible state to be deductively closed or to be consistent, even if none of its component sentences express non-trivial reasons. We introduce such general restrictions on the set of admissible states because interpreted components alone cannot express all interesting restrictions. Actually, nonemptiness of admissible states is the only restriction inexpressible by components alone, for though $\emptyset$ is always component-admissible (it having no elements to rule otherwise), we can always replace $\mathcal{I}$ by $\mathcal{I}'$, defined so that for all $d \in \mathcal{D}$, $\mathcal{I}'(d) = \mathscr{S}$, in which case $\mathcal{Q} = \mathscr{S} \cup \{\emptyset\}$.

In many cases of interest, as in the examples of deductive closure, consistency, and nonemptiness above, it is possible to decompose the specification of $\mathscr{S}$ into the component restrictions and a general restriction on all sets. That is, one way of specifying $\mathscr{S} \subseteq \mathcal{Q}$ is to stipulate a general restriction on states $\mathcal{R} \subseteq \mathbf{P}\,\mathcal{D}$ and define

$$\mathscr{S} = \mathcal{Q} \cap \mathcal{R} = \{S \in \mathcal{R} \mid S \in \bigcap_{d \in S} \mathcal{I}(d)\}.$$

Since $\mathscr{S} \subseteq \mathcal{Q}$, we can always take $\mathcal{R} = \mathscr{S}$ when there is no independently interesting definition of $\mathcal{R}$. If $\mathscr{S} = \mathcal{R}$, then the component interpretations add nothing to the general restriction. In particular, this circumstance holds whenever all component interpretations are trivial. On the other hand, if $\mathcal{R} = \mathbf{P}\,\mathcal{D}$, then there are no general restrictions, and any restrictions on admissible states must be explicit in the states themselves. This circumstance recalls the current efforts in artificial intelligence towards constructing completely "self-descriptive" machines.[12] In the theories to follow, we will define $\mathscr{S}$ either directly or in terms of $\mathcal{R}$ as convenient.

§**13.** While the notion of admissible state captures an idea of states which specify their own structure, it says nothing about any form of inference. We introduce the idea of *admissible extension* as a formalization of the sets of conclusions or reasoned assumptions permitted within the structure of the agent's states. Just as the specification of admissible states involved both "local" ($\mathcal{I}$) and "global" ($\mathscr{S}$ or $\mathcal{R}$) restrictions, so also does the definition of admissible extension.

If $S \subseteq \mathcal{D}$ we define $\mathit{Exts}(S)$, the set of *extensions of* or *states extending* S, to be the admissible states including S as a subset, or formally,

$$\mathit{Exts}(S) = \{E \in \mathscr{S} \mid S \subseteq E\}.$$

If $E \in \mathit{Exts}(S)$, we also write $S \lhd E$. Extensions are defined in the same way for all theories of reasoned assumptions.

Just as the "psycho-logic" of mental states interprets (via $\mathcal{I}$) each state component as a restriction on the states in which it can occur, we also interpret components as restrictions on the ways they can be derived or occur in admissible extensions. That is, we assume a function $\mathcal{J} : \mathcal{D} \times \mathbf{P}\,\mathcal{D} \to \mathbf{P}\,\mathbf{P}\,\mathcal{D}$ that interprets each state component to find the extensions it sanctions for various sets of components. We define $\mathit{QExts}(S)$, the *component-admissible extensions* of a set $S \subseteq \mathcal{D}$, by

$$\mathit{QExts}(S) = \{E \in \mathit{Exts}(S) \mid E \in \bigcap_{d \in E} \mathcal{J}(d, S)\}.$$

That is, E is a component-admissible extension of S just in case each element of E approves (via $\mathcal{J}$) of the way it occurs in E relative to S. The admissible extensions $\mathit{AExts}(S)$ are stipulated as a subset of the component-admissible extensions, or formally, $\mathit{AExts}(S) \subseteq \mathit{QExts}(S) \subseteq \mathit{Exts}(S)$. If $E \in \mathit{AExts}(S)$, we also write $S \lhd\!\!\lhd E$.

Putting all these definitions together, we say that each choice of $(\mathcal{D}, \mathcal{I}, \mathscr{S}, \mathcal{J}, \lhd\!\!\lhd)$ (or alternatively, each choice of $(\mathcal{D}, \mathcal{I}, \mathcal{R}, \mathcal{J}, \lhd\!\!\lhd)$) describes the states and inferential structure of an *autological agent*. This appelation is meant to recall the ways the agent's constituents talk about their roles in the "logic" of the agent's states.

[12] [MINSKY 1965], [DOYLE 1980], [WEYHRAUCH 1980], [FRISCH AND ALLEN 1982], [SMITH 1982]

§**14.** Merely having structures for admissible states and admissible extensions does not in itself support use of the term "logic," but we introduce a notion of psychological entailment to excuse this abuse of standard terminology. Let $(\mathcal{D}, \mathcal{I}, \mathscr{S}, \mathcal{J}, \lhd)$ describe an autological agent. If $A, B, S \subseteq \mathcal{D}$, we say that A *psychologically entails* B *in* S (within the agent's psychology), written $A \Vdash_S B$, iff $B \subseteq E$ whenever $A \subseteq E$ and $E \in AExts(S)$.

(14.1) THEOREM. *$A \Vdash_S B$ for every $A, B \subseteq \mathcal{D}$ iff $AExts(S) = \emptyset$ or $AExts(S) = \{\mathcal{D}\}$.*

PROOF. Clearly, if $AExts(S)$ is $\emptyset$ or $\{\mathcal{D}\}$, then $A \Vdash_S B$ for every $A, B \subseteq \mathcal{D}$, so suppose $A \Vdash_S B$ for every $A, B \subseteq \mathcal{D}$ and $AExts(S) \neq \emptyset$. Let $E \in AExts(S)$. Then by hypothesis, $\emptyset \Vdash_S \mathcal{D}$, and since $\emptyset \subseteq E$, we must have $\mathcal{D} \subseteq E$, hence $\mathcal{D} = E$. ∎

A more specific notion plays a prominent role in the subsequent development where we say that $d \in \mathcal{D}$ is *inevitable* in S to mean that $d \in E$ whenever $E \in AExts(S)$, that $\emptyset \Vdash_S \{d\}$.

We introduced the notion of admissible extension to capture the notion of inference within the agent's psychology, but unfortunately, the correspondence between inference and entailment so important in mathematical logic does not obtain in the general psychology. We say that B *is arguable from* A *in* S, written $A \mathrel{|\!\sim}_S B$, iff there is some $E \in AExts(S \cup A)$ such that $B \subseteq E$.

(14.2) THEOREM. *If $A \mathrel{|\!\sim}_S B$ for every $A, B \subseteq \mathcal{D}$, then $S \lhd \mathcal{D}$.*

PROOF. Suppose $A \mathrel{|\!\sim}_S B$ for every $A, B \subseteq \mathcal{D}$. Then in particular, $\emptyset \mathrel{|\!\sim}_S \mathcal{D}$, so there is some $E \in AExts(S)$ such that $\mathcal{D} \subseteq E$, hence $\mathcal{D} = E$ and $S \lhd \mathcal{D}$. ∎

We say that $d \in \mathcal{D}$ is *arguable* in S iff there is some $E \in AExts(S)$ such that $d \in E$, in other words, $\emptyset \mathrel{|\!\sim}_S \{d\}$.

(14.3) COROLLARY. *If $A \subseteq S$, then $A \Vdash_S B$ iff $\emptyset \Vdash_S B$, and $A \mathrel{|\!\sim}_S B$ iff $\emptyset \mathrel{|\!\sim}_S B$.*

The divergence between arguability and inevitability is seen more clearly in the particular subsequent theories. Indeed, the point of theories of reasoned assumptions is to set out the coherent sets of assumptions sanctioned by some set of reasons, and when there are several possible coherent sets of assumptions one has $A \mathrel{|\!\sim}_S B$ but not $A \Vdash_S B$, a reflection of the intended "unsoundness" of this sort of psychological inference.

§**15.** Even among the "unsound" conclusions sanctioned as reasoned assumptions, some conclusions are more sound than others. If $E, E' \in AExts(S)$ are such that E is a proper subset of E', then those conclusions in E' but not in E are, in a sense, less sound than need be. Expedience might force jumping to conclusions, but it need not force profligacy. To avoid unnecessary unsoundness, we introduce the notion of *strict arguability*, arguability in minimal admissible extensions.

(15.1) DEFINITION. *If X is a set of sets, the minimization of X, μX, is given by $\mu X = \{x \in X \mid \forall y \in X\ y \subseteq x \supset x \subseteq y\}$. If f is a set function $f : X \to \mathbf{P}Y$, then the minimization of f is a function $\mu f : X \to \mathbf{P}Y$ such that for each $x \in X$*

$$(\mu f)(x) = \mu(f(x)) = \{y \in f(x) \mid\ \forall y' \in f(x) \quad y' \subseteq y \supset y \subseteq y'\}.$$

Take care to note that μ finds the minimal elements of sets, not the minimums of sets, so that $\mu f(x) = \emptyset$ if $f(x) = \emptyset$. Note also that μX can be empty even when X is not if X contains an infinite descending chain of sets. This will not happen in most of the theories we consider. One set function is $AExts : \mathbf{P}\mathcal{D} \to \mathbf{P}\mathscr{S}$, and its minimization under set inclusion, written $\mu AExts$, is defined for all $S \subseteq \mathcal{D}$ by

$$\mu AExts(S) = \{E \in AExts(S) \mid\ \forall E' \in AExts(S) \quad E' \subseteq E \supset E \subseteq E'\}.$$

While this definition of minimization is specific to set inclusion, we elsewhere treat these ideas in a more general setting by replacing set inclusion with more specific notions of relative information content. In that treatment, μ

means minimization with respect to a quasi-order $\sqsubseteq$ on $\mathbf{P}\,\mathcal{D}$ refined by $\subseteq$, that is, an order $\sqsubseteq$ such that $x \sqsubseteq y$ implies $x \subseteq y$ but not necessarily vice versa.

With the definition of minimization, we say that B is *strictly arguable from A in S*, written $A\mu\!\mid\!\sim_S B$, iff there is some $E \in \mu AExts(S \cup A)$ such that $B \subseteq E$. In this way, strict arguability corresponds to arguability with as few assumptions as possible. Note that if $\mu AExts = AExts$ the two versions of arguability coincide.

Corresponding to the notion of strict arguability in an agent, we also have the notion of minimal psychological entailment, psychological entailment in minimal admissible extensions. Specifically, we say A *minimally psychologically entails B in S*, written $A\mu \,\|\!\!=_S B$, iff $B \subseteq E$ whenever $A \subseteq E$ and $E \in \mu AExts(S)$. Since $\mu AExts(S) \subseteq AExts(S)$, it is clear that $A\mu \,\|\!\!=_S B$ whenever $A \,\|\!\!=_S B$. Unlike strict arguability, minimal psychological entailment does not play a significant role in the following theories of reasoned assumptions. It is discussed in more detail in §51 along with the logical notion of circumscription.

§**16.** Several other inferential relationships are also worth naming. Let $d, e \in \mathcal{D}$ and $S \subseteq \mathcal{D}$. We say d and e are *cotenable* in S if there is some $E \in AExts(S)$ such that $d, e \in E$, or put another way, if $\emptyset \mid\!\sim_S \{d, e\}$. We say S is *coherent* if $AExts(S) \neq \emptyset$, and is *incoherent* otherwise. We say that d is *assumable* in S if $S \cup \{d\}$ is coherent, and *realizable* in S if some $S' \supset S \cup \{d\}$ is coherent. Clearly, S is coherent iff $\emptyset \mid\!\sim_S S$, d is assumable iff $\{d\} \mid\!\sim_S \{d\}$, and d is realizable iff for some $A \subseteq \mathcal{D}$, $A \cup \{d\} \mid\!\sim_S \{d\}$. It is also easy to see that if d and e are cotenable in S, each is arguable in S; that if S is coherent and d is inevitable in S, d is arguable in S; that if S is coherent and both d and e are inevitable in S, d and e are cotenable in S; and that if S is coherent and $A, B \subseteq S$, then $A \mid\!\sim_S B$.

If $\mathcal{D}$ has the structure of the set of sentences of an ordinary logical language, there is a 1-1 function $\neg : \mathcal{D} \to \mathcal{D}$ taking elements to negations. For this important special class of domains, we introduce the following terms. We say d is *doubtless* in S iff $\neg d$ is not arguable in S. Similarly, d is *conceivable* in S iff $\neg d$ is not inevitable in S. Arguability and doubtlessness are thus dual notions, as are inevitability and conceivability. We say d is *decided* by S iff for every $E \in AExts(S)$ either $d \in E$ or $\neg d \in E$, and that S is *ambivalent* about d iff d is not decided by S. We say S is *arguably consistent* if there is some $E \in AExts(S)$ such that for every $d \in \mathcal{D}$, either $d \notin E$ or $\neg d \notin E$, and *arguably inconsistent* if for some d and E, $d, \neg d \in E \in AExts(S)$. We say S is *inevitably consistent* (or simply *consistent*) if S is coherent and for every $E \in AExts(S)$ and $d \in \mathcal{D}$, either $d \notin E$ or $\neg d \notin E$. S is *inevitably inconsistent* if S is coherent but there is some d such that for every $E \in AExts(S)$, $d \in E$ and $\neg d \in E$. It is clear that if S is consistent, S is arguably consistent; that if d and $\neg d$ are cotenable in S, S is inconsistent; that if d is doubtless in S yet decided by S, d is inevitable in S; and that if d is inevitable in S and e is arguable in S, d and e are cotenable in S. Note that these notions are weaker than the usual notions of consistency, for we have made no assumptions about iterated negations, that is we allow $d \in S$ but $\neg\neg d \notin S$. If $\mathcal{D}$ has the structure of the Lindenbaum algebra of sentences of a logical language, then $\neg\neg$ is the identity. In this case it is clear that d is decided by S iff $\neg d$ is decided by S, and that S is ambivalent about d iff $\neg d$ is not decided by S. We pursue these more familiar notions later under the topic *Logical theories*.

§**17.** As is usual with formal systems, we can construct bigger autological agents out of smaller ones. Two basic constructions are sum and product agents. Let $(\mathcal{D}_1, \mathcal{I}_1, \mathscr{S}_1, \mathcal{J}_1, \lhd_1)$ and $(\mathcal{D}_2, \mathcal{I}_2, \mathscr{S}_2, \mathcal{J}_2, \lhd_2)$ describe two autological agents. The sum of these is an agent $(\mathcal{D}, \mathcal{I}, \mathscr{S}, \mathcal{J}, \lhd)$ such that

(1) $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$,

(2) $\mathcal{I}(d) = \{S \subseteq \mathcal{D} \mid d \in \mathcal{D}_1 \supset [S \cap \mathcal{D}_1 \in \mathcal{I}_1(d)] \ \wedge \ d \in \mathcal{D}_2 \supset [S \cap \mathcal{D}_2 \in \mathcal{I}_2(d)]\}$

(3) $\mathscr{S} = \{S \subseteq \mathcal{D} \mid S \cap \mathcal{D}_1 \in \mathscr{S}_1 \ \wedge \ S \cap \mathcal{D}_2 \in \mathscr{S}_2\}$

(4) $\mathcal{J}(d, S) = \{E \mid d \in \mathcal{D}_1 \supset [E \cap \mathcal{D}_1 \in \mathcal{J}_1(d, S \cap \mathcal{D}_1)] \ \wedge \ d \in \mathcal{D}_2 \supset [E \cap \mathcal{D}_2 \in \mathcal{J}_2(d, S \cap \mathcal{D}_2)]\}$

(5) $\lhd = \{(S, S') \mid S \cap \mathcal{D}_1 \lhd_1 S' \cap \mathcal{D}_1 \ \wedge \ S \cap \mathcal{D}_2 \lhd_2 S' \cap \mathcal{D}_2\}$.

The product of the two agents is an agent such that

(1) $\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2$,

(2) $\mathcal{I}((d_1, d_2)) = \{S \subseteq \mathcal{D} \mid \exists S_1 \in \mathcal{I}_1(d_1)\ \exists S_2 \in \mathcal{I}_2(d_2) \quad S = S_1 \times S_2\}$,

(3) $\mathscr{S} = \{S \subseteq \mathcal{D} \mid \exists S_1 \in \mathscr{S}_1\ \exists S_2 \in \mathscr{S}_2 \quad S = S_1 \times S_2\}$,

(4) $\mathcal{J}((d_1, d_2), S) = \{E \mid \exists S_1, E_1 \subseteq \mathcal{D}_1\ \exists S_2, E_2 \subseteq \mathcal{D}_2$
$S = S_1 \times S_2 \ \wedge \ E = E_1 \times E_2 \ \wedge$
$E_1 \in \mathcal{J}_1(d_1, S_1) \ \wedge \ E_2 \in \mathcal{J}_2(d_2, S_2)\}$,

(5) $\lhd = \{(S, S') \mid \exists S_1, S_1' \subseteq \mathcal{D}_1\ \exists S_2, S_2' \subseteq \mathcal{D}_2 \quad S = S_1 \times S_2 \ \wedge \ S' = S_1' \times S_2' \ \wedge$
$S_1 \lhd_1 S_2 \ \wedge \ S_1' \lhd_2 S_2'\}$.

The sum construction appears quite frequently in artificial intelligence systems, which often divide into independent databases operating with different inferential schemes.

§**18.** The first theory of reasoned assumptions we consider is the theory of *simple reasons*. Simple reasons are so named because they involve what seem to be the weakest useful notion of reason, one on which all the following theories elaborate. This theory makes no assumptions about $\mathcal{D}$, has no general restrictions on states, and interprets all components as (possibly trivial) "simple reasons," in contrast to the "defeasible reasons" defined in the next section. We split the definition into two parts due to some intervening subsidiary definitions.

(18.1) DEFINITION (SIMPLE REASONS I). *An agent's use of simple reasons is characterized by* $(\mathcal{D}, \mathcal{I}, \mathscr{S}, \mathcal{J})$*, where*

(i) $\mathcal{D}$ *is a set,*

(ii) For each $d \in \mathcal{D}$*, there are sets* $A, B, C \subseteq \mathcal{D}$ *such that*

$$\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\sim C = \{S \subseteq \mathcal{D} \mid A \subseteq S \subseteq B^c \supset C \subseteq S\},$$

(iii) $\mathscr{S} = \mathcal{Q}$*,*

(iv) For each $d \in \mathcal{D}$ *and* $S \subseteq \mathcal{D}$*,*

$$\mathcal{J}(d, S) = \{E \mid d \in E \supset [d \in S \quad \vee \quad \exists e \in E\ \exists A, B, C \subseteq \mathcal{D} \\ \mathcal{I}(e) = A \setminus\setminus B \mid\!\!\sim C\ \wedge\ A \subseteq E \subseteq B^c\ \wedge\ d \in C]\}.$$

Recalling our earlier notation, we abbreviate the subsets of $\mathbf{P}\,\mathcal{D}$ corresponding to reason interpretations as expressions of the form $A \setminus\setminus B \mid\!\!\sim C$, where $A, B, C \subseteq \mathcal{D}$. The definition of $\mathcal{I}$ interprets reasons so that the "conclusions" C must be held if the "antecedents" A are held and none of the "qualifiers" B are held. Note that expressions like $A \setminus\setminus B \mid\!\!\sim C$ are part of the metalanguage in which we discuss the agent. The agent's language of thought, if any, need not express reasons in the same way. The notation $A \setminus\setminus B \mid\!\!\sim C$ also allows us to speak of the "same" reason even when we extend $\mathcal{D}$ to a larger domain, since if $A, B, C \subseteq \mathcal{D}$ and $\mathcal{D} \subseteq \mathcal{D}'$, we also have $A, B, C \subseteq \mathcal{D}'$. This property greatly simplifies mechanizations of agents based on simple reasons expressed in this way, since the domain of state components can be extended indefinitely without necessitating changes in previously expressed reasons. Finally, observe that the expression $\emptyset \setminus\setminus \emptyset \mid\!\!\sim \emptyset$ means the trivial interpretation, the set $\mathbf{P}\,\mathcal{D}$.

(18.2) DEFINITION. *A reason with interpretation* $A \setminus\setminus B \mid\!\!\sim C$ *is said to be valid in* S *iff its antecedent conditions hold, that is, iff* $A \subseteq S \subseteq B^c$.

In this way, component-admissible extensions are those extensions which locally appear "grounded" since each of the elements is in the initial set or is a conclusion of a valid reason. We base the general restriction giving admissible extensions on a notion of globally grounded extensions.

(18.3) DEFINITION. *A finite reason is an element* $d \in \mathcal{D}$ *such that there are finite subsets* $A, B, C \subseteq \mathcal{D}$ *with* $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\sim C$.

(18.4) DEFINITION. E *is a (finitely) grounded extension of* S *iff* $S \lhd E$ *and for each* $e \in E$ *there is a (finite) grounding set* $G \subseteq E$ *and a well-ordering* $<_G$ *of* G *such that* $e \in G$ *and whenever* $d \in G$*, either (1)* $d \in S$ *or (2) there is some (finite) reason* $f \in G$ *and sets* $A, B, C \subseteq \mathcal{D}$ *such that* $\mathcal{I}(f) = A \setminus\setminus B \mid\!\!\sim C$*,* $A \subseteq G$*,* $E \subseteq B^c$*,* $d \in C$*, and* $A <_G f <_G d$*. For each* $S \subseteq \mathcal{D}$*,* $GExts(S)$ *denotes the set of grounded extensions of* S*, and* $FGExts(S)$ *the set of finitely grounded extensions of* S.

(18.5) COROLLARY. *If* E *is a grounded extension of* S*,* $e \in E$*,* G *is a grounding set of* e*, and* $d \leq_G e$*, then* $\{g \in G \mid g \leq_G d\}$ *is a grounding set for* d *in* E*, as is* G *itself.*

Grounding sets bear a remarkable similarity to open neighborhoods in a topological space, but I have not been able to make them into such since the intersection of two grounding sets need not be a grounding set.

(18.6) COROLLARY. $FGExts \subseteq GExts$.

(18.7) THEOREM. $GExts \subseteq QExts$.

PROOF. The claim is just that grounded extensions also appear locally grounded. Suppose $E \in GExts(S)$ and $e \in E$. Then if $e \notin S$, there is a grounding set G for e. But since $e \in G$ and $e \notin S$, there is a valid reason supporting e in G, and hence in E as well. ∎

(18.8) DEFINITION (SIMPLE REASONS II). *Grounded, finitely grounded, and locally grounded simple reasons agents are characterized respectively by $AExts = GExts$, $AExts = FGExts$, and $AExts = QExts$.*

For the sake of simplicity and practical relevance we normally discuss finitely grounded simple reasons agents, take $AExts = FGExts$, and explicitly state whenever we consider general grounded or locally grounded agents.

EXAMPLES. We adopt the convention that if $\neg : \mathcal{D} \to \mathcal{D}$ and $A \subseteq \mathcal{D}$ we write $\neg A$ to mean $\{\neg a \mid a \in A\}$, and if $S \subseteq \mathcal{D}$, we write $\mathcal{I}^*(S)$ to mean $\{\mathcal{I}(d) \mid d \in S\}$. In all the examples, we usually assume A, B, C, D and their negations to be subsets of $\mathcal{D}$ composed of trivially interpreted elements, so if $a \in A$, then $\mathcal{I}(a) = \mathcal{I}(\neg a) = \emptyset \setminus\!\setminus \emptyset \mid\!\!\sim \emptyset$. Comments, if any, follow the example to which they refer.

(18.9) $\mathcal{I}^*(S) = \{\emptyset \setminus\!\setminus \emptyset \mid\!\!\sim \emptyset\}, \qquad AExts(S) = \{S\}$

If S has only trivially interpreted elements, it is its own admissible extension.

(18.10) $\mathcal{I}^*(S) = \{\emptyset \setminus\!\setminus \emptyset \mid\!\!\sim A\}, \qquad AExts(S) = \{S \cup A\}$

This sort of reason can be viewed as setting out unqualified premises.

(18.11) $\mathcal{I}^*(S) = \{A \setminus\!\setminus \emptyset \mid\!\!\sim A\}, \qquad AExts(S) = \{S\}$

Note that $S \cup A$ is a component-admissible extension of S, but neither a grounded component-admissible extension of S, nor a minimal component-admissible extension of S, nor a minimal extension of S.

(18.12) $\mathcal{I}^*(S) = \{\emptyset \setminus\!\setminus \neg A \mid\!\!\sim A\}, \qquad AExts(S) = \{S \cup A\}$

This sort of reason is commonly called a *normal default*; a "default" since it sanctions an inference in which one draws a conclusion because one has no valid reason for drawing the opposite conclusion, and "normal" since this sort of default is so commonly useful in artificial intelligence. Note that $S \cup \neg A$ is also a minimal extension of S, but is not component-admissible.

(18.13) $\mathcal{I}^*(S) = \{\emptyset \setminus\!\setminus A \mid\!\!\sim A\}, \qquad AExts(S) = \emptyset$

Note that while $S \cup A$ is an admissible state, moreover a minimal extension of S, it is not component-admissible.

(18.14) $\mathcal{I}^*(A) = \{\emptyset \setminus\!\setminus C \mid\!\!\sim B\}, \quad \mathcal{I}^*(B) = \{\emptyset \setminus\!\setminus \emptyset \mid\!\!\sim \emptyset\},$
$\mathcal{I}^*(C) = \{\emptyset \setminus\!\setminus \emptyset \mid\!\!\sim D\}, \quad \mathcal{I}^*(D) = \{\emptyset \setminus\!\setminus \emptyset \mid\!\!\sim C\}$
$AExts(A) = \{A \cup B\}$

Note that $A \cup C \cup D$ is a minimal component-admissible extension of A, but not a grounded extension of A.

(18.15) $\mathcal{I}^*(S) = \{\emptyset \setminus\!\setminus \neg A \mid\!\!\sim A, \quad \emptyset \setminus\!\setminus A \mid\!\!\sim \neg A\}, \qquad AExts(S) = \{S \cup A, \quad S \cup \neg A\}$

This example exhibits the multiple admissible extensions of the motivating examples of §6.

(18.16) $\mathcal{I}^*(S) = \{\emptyset \setminus\!\setminus A \mid\!\!\sim A, \quad \emptyset \setminus\!\setminus \emptyset \mid\!\!\sim A\}, \qquad AExts(S) = \{S \cup A\}$

In contrast to 18.13, the second reason supports A regardless of the incoherence of the first reason.

(18.17) $\mathcal{I}^*(S) = \{\emptyset \setminus\!\setminus A \mid\!\!\sim A, \quad \emptyset \setminus\!\setminus \emptyset \mid\!\!\sim \neg A\}, \qquad AExts(S) = \emptyset$

Since we have provided no connection between the interpretations of A and $\neg A$, this example has the same basic structure as 18.13.

(18.18) $\mathcal{I}^*(S) = \{\emptyset \setminus\!\setminus \emptyset \mid\!\!\sim A, \quad \emptyset \setminus\!\setminus \emptyset \mid\!\!\sim \neg A\}, \qquad AExts(S) = \{S \cup A \cup \neg A\}$

Since the agent has not been given logical structure, we draw no further conclusions from this inconsistent extension.

(18.19) $\mathcal{I}^*(S) = \{\emptyset \setminus\!\setminus \neg A \mid\!\!\sim A, \quad \emptyset \setminus\!\setminus \neg B \mid\!\!\sim B, \quad A \setminus\!\setminus \emptyset \mid\!\!\sim \neg B, \quad B \setminus\!\setminus \emptyset \mid\!\!\sim \neg A\},$
$AExts(S) = \{S \cup A \cup \neg B, \quad S \cup \neg A \cup B\}$

(18.20) $\mathcal{I}^*(S) = \{\emptyset \setminus\!\setminus \neg A \mid\!\!\sim A, \quad \emptyset \setminus\!\setminus \neg B \mid\!\!\sim B, \quad \emptyset \setminus\!\setminus \neg C \mid\!\!\sim C, \quad A \setminus\!\setminus \emptyset \mid\!\!\sim \neg B,$
$B \setminus\!\setminus \emptyset \mid\!\!\sim \neg C, \quad C \setminus\!\setminus \emptyset \mid\!\!\sim \neg A\},$
$AExts(S) = \emptyset$

(18.21) $\mathcal{I}^*(S) = \{\emptyset \setminus\!\setminus \neg A \mid\!\!\sim A, \quad \emptyset \setminus\!\setminus \neg B \mid\!\!\sim B, \quad \emptyset \setminus\!\setminus \neg C \mid\!\!\sim C, \quad \emptyset \setminus\!\setminus \neg D \mid\!\!\sim D,$
$A \setminus\!\setminus \emptyset \mid\!\!\sim \neg B, \quad B \setminus\!\setminus \emptyset \mid\!\!\sim \neg C, \quad C \setminus\!\setminus \emptyset \mid\!\!\sim \neg D, \quad D \setminus\!\setminus \emptyset \mid\!\!\sim \neg A\},$
$AExts(S) = \{S \cup A \cup \neg B \cup C \cup \neg D, \qquad S \cup \neg A \cup B \cup \neg C \cup D\}$

Note how the parity of these cycles (even in 18.19 and 18.21, odd in 18.20) affects the existence of admissible

extensions.

(18.22) $\mathcal{I}^*(S) = \{\emptyset \setminus\setminus B \mid\!\!\!\vdash A, \quad \emptyset \setminus\setminus \emptyset \mid\!\!\!\vdash \neg A\}, \qquad AExts(S) = \{S \cup A \cup \neg A\}$

Note that contrary conclusions cannot "push backwards" through reasons to support qualifiers.

(18.23) Suppose $\{d_i\}_{i=0}^{\infty}$ is a set of distinct elements of $\mathcal{D}$ such that for each $i \geq 0$, $\mathcal{I}(d_i) = \emptyset \setminus\setminus \emptyset \mid\!\!\!\vdash \{d_{i+1}\}$. Then $AExts(\{d_i\}) = \{\{d_j \mid j \geq i\}\}$. Even though each set $\{d_i\}$ is finite, its admissible extension is infinite.

(18.24) Let $D = \{d_i\}_{i=0}^{\infty}$ and $\mathcal{D} = D \oplus \{e, f\}$, where for each $i \geq 0$, $\mathcal{I}(d_i) = \emptyset \setminus\setminus \emptyset \mid\!\!\!\vdash \{d_{i+1}\}$, and where $\mathcal{I}(e) = D \setminus\setminus \emptyset \mid\!\!\!\vdash \{f\}$ and $\mathcal{I}(f) = \emptyset \setminus\setminus \emptyset \mid\!\!\!\vdash \emptyset$. Let $S = \{d_0, e\}$. Then $GExts(S) = \{\mathcal{D}\}$, but $FGExts(S) = \emptyset$ since no finite argument for f exists.

I apologize for several little white lies among these examples. These stem from the alternative expressions that some reason interpretations allow. For example, the expression $A \setminus\setminus \emptyset \mid\!\!\!\vdash A$ means the same set as does $\emptyset \setminus\setminus \emptyset \mid\!\!\!\vdash \emptyset$, namely $\mathbf{P}\,\mathcal{D}$. Further, the expression $\emptyset \setminus\setminus A \mid\!\!\!\vdash A$ means the same as does $\emptyset \setminus\setminus \emptyset \mid\!\!\!\vdash A$. Thus it is wrong to call $\{\emptyset \setminus\setminus A \mid\!\!\!\vdash A\}$ incoherent yet to call $\{\emptyset \setminus\setminus \emptyset \mid\!\!\!\vdash A\}$ coherent. We expressed the examples as simply as possible for the sake of comprehension, but these difficulties can always be avoided by slightly more complex statements of examples. For example, if $\mathcal{I}(a) = \emptyset \setminus\setminus \{b\} \mid\!\!\!\vdash \{c\}$, $\mathcal{I}(b) = \mathcal{I}(c) = \emptyset \setminus\setminus \emptyset \mid\!\!\!\vdash \emptyset$, and $\mathcal{I}(d) = \{c\} \setminus\setminus \emptyset \mid\!\!\!\vdash \{b\}$, then the set $\{a, d\}$ is incoherent, while $\{a, d, c\}$ is coherent.

In spite of their simplicity and practical significance, very little is known about the properties of simple reason agents. The few results which follow are merely indicative of the questions that remain to be answered. We first examine some alternate characterizations of the notion of admissible extension.

(18.25) DEFINITION. *Let $S, E \subseteq \mathcal{D}$. Then $\langle \Lambda_\alpha \rangle$ (α an ordinal), the levels from S in E, are defined for all ordinals by*

$$\Lambda_0(S, E) = S,$$

$$\Lambda_{\alpha+1}(S, E) = \Lambda_\alpha \cup \bigcup \{C \subseteq \mathcal{D} \quad | \quad \exists e \in \Lambda_\alpha(S, E)\ \exists A, B \subseteq \mathcal{D}$$
$$\mathcal{I}(e) = A \setminus\setminus B \mid\!\!\!\vdash C \ \wedge\ A \subseteq \Lambda_\alpha(S, E) \ \wedge\ E \subseteq B^c\},$$

and for limit ordinals λ,

$$\Lambda_\lambda(S, E) = \bigcup_{\alpha < \lambda} \Lambda_\alpha(S, E).$$

We also define $\Lambda(S, E) = \bigcup_\alpha \Lambda_\alpha(S, E)$ to be the sum of all levels. When no confusion is possible, we sometimes abbreviate $\Lambda_\alpha(S, E)$ by Λ_α and $\Lambda(S, E)$ by Λ.

(18.26) COROLLARY. *If $\alpha \leq \beta$, then $S \subseteq \Lambda_\alpha(S, E) \subseteq \Lambda_\beta(S, E) \subseteq \Lambda(S, E)$.*

(18.27) COROLLARY. *If $\Lambda_\alpha(S, E) = \Lambda_{\alpha+1}(S, E)$, then $\Lambda_\alpha(S, E) = \Lambda(S, E)$.*

(18.28) THEOREM. *If α is the cardinal number $|\mathcal{D}| + 1$, then $\Lambda(S, E) = \Lambda_\alpha(S, E)$.*

PROOF. Since $\mathcal{D}$ has fewer than α elements, it must be that for some $\beta + 1 \leq \alpha$ no new element is introduced in $\Lambda_{\beta+1}$, in other words, $\Lambda_\beta = \Lambda_{\beta+1}$. But then $\Lambda_\beta = \Lambda$, and since $\Lambda_\beta \subseteq \Lambda_\alpha \subseteq \Lambda$, we have $\Lambda = \Lambda_\alpha$. ∎

(18.29) DEFINITION. *If $e \in \Lambda(S, E)$, the rank of e in $\Lambda(S, E)$ is the least ordinal α such that $e \in \Lambda_\alpha(S, E)$. If $A \subseteq \Lambda(S, E)$, the rank of A in $\Lambda(S, E)$ is the least ordinal not less than the rank of any element of A.*

(18.30) LEMMA. *If $S \lhd E$, then $\Lambda(S, E) \subseteq E$.*

PROOF. Let $S \lhd E$. Clearly $\Lambda_0 \subseteq E$, so assume $\Lambda_\beta \subseteq E$ for each $\beta < \alpha$. If α is a limit ordinal, then by definition $\Lambda_\alpha \subseteq E$. If α is a successor ordinal, say $\alpha = \beta + 1$, let $e \in \Lambda_\alpha$. If $e \in S$, then $e \in E$, and if $e \notin S$ there is a $d \in \Lambda_\beta$ with $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\!\vdash C$, $A \subseteq \Lambda_\beta$, $E \subseteq B^c$, and $e \in C$. Since E is admissible, this means $C \subseteq E$, so $e \in E$, hence $\Lambda_\alpha \subseteq E$. Thus $\Lambda \subseteq E$. ∎

(18.31) THEOREM (STRATIFICATION). *If $E \in GExts(S)$ then $\Lambda(S, E) = E$.*

PROOF. Suppose $E \in GExts(S)$. Since $S \lhd E$, by the preceding lemma we have $\Lambda \subseteq E$. To see that $E \subseteq \Lambda$, suppose $e \in E$. Since E is a grounded extension of S, there is a grounding set $G \subseteq E$ for e from S in E. We show $G \subseteq \Lambda$ by $<_G$-induction. Let $f \in G$ have no predecessors in $<_G$. Clearly f is the minimum of G, and by definition of G, we must have $f \in S$, hence $f \in \Lambda$. Now suppose that $f \in G$ and for each $d <_G f$, either $d \in S$ or there is a grounding subargument $G' \subseteq G$ for d. If $f \in S$, then $f \in \Lambda$, and if $f \notin S$, there is a $d \in G$ such that $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\!\vdash C$, $A <_G d <_G f$, $E \subseteq B^c$, and $f \in C$. By the inductive hypothesis, $A \subseteq \Lambda$ and $d \in \Lambda$, so there is some ordinal α such that $A \subseteq \Lambda_\alpha$ and $d \in \Lambda_\alpha$. But then by construction $C \subseteq \Lambda_{\alpha+1}$, so $f \in \Lambda$. Hence $E \subseteq \Lambda$, so $E = \Lambda(S, E)$. ∎

(18.32) COROLLARY. *If $E \in GExts(S)$ and $\alpha = |E| + 1$, then $E = \Lambda_\alpha(S, E)$.*

(18.33) COROLLARY. *If $E \in GExts(S)$ and α is its rank, then $E = \Lambda_\alpha(S, E)$.*

(18.34) COROLLARY. *If $E \in FGExts(S)$, then $E = \Lambda_\omega(S, E)$.*

PROOF. Let $E \in FGExts(S)$ and $e \in E$. Since e has a finite grounding set G, the rank of e is at most $|G|$, hence $e \in \Lambda_\omega(S, E)$. Thus $E \subseteq \Lambda_\omega(S, E)$, so by Lemma 18.30, $E = \Lambda_\omega(S, E)$. ∎

(18.35) THEOREM (FIXED POINT). *If $E = \Lambda(S, E)$, then $E \in GExts(S)$.*

PROOF. Suppose $\Lambda = E$. Since $S \subseteq \Lambda$, $S \subseteq E$. Let $e \in E$ with $\mathcal{I}(e) = A \setminus\setminus B \mid\!\!\!\vdash C$, and suppose $A \subseteq E$. Then there is an ordinal α such that $e \in \Lambda_\alpha$ and $A \subseteq \Lambda_\alpha$, so by construction if $E \subseteq B^c$ as well, then $C \subseteq \Lambda_{\alpha+1} \subseteq E$. Thus E is admissible. We prove E is a grounded extension of S by induction on rank. Specifically, we prove that each element of E has a rank-preserving grounding set, a set $G \subseteq E$ such that *rank*$(a) \leq$ *rank*(b) whenever $a \leq_G b$. Let $e \in E$ have rank α. If $\alpha = 0$, then $e \in S$ and we are done since $\{e\}$ is a rank-preserving grounding argument for e from S in E. Now assume that $\alpha > 0$ and all elements of rank less than α have rank-preserving grounding arguments. By construction, there is some $\beta < \alpha$ and $d \in \Lambda_\beta$ such that $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\!\vdash C$, $A \subseteq \Lambda_\beta$, $E \subseteq B^c$, and $e \in C$. Then by inductive hypothesis each element of $\{d\} \cup A$ has a rank-preserving grounding argument, so merge these arguments preserving rank-order, and add e to the end, so producing a rank-preserving grounding argument for e. Thus $E \in GExts(S)$. ∎

(18.36) COROLLARY. *$E \in GExts(S)$ iff $E = \Lambda(S, E)$.*

(18.37) THEOREM. *If every reason in $\mathcal{D}$ is finite, then $E \in FGExts(S)$ iff $E = \Lambda_\omega(S, E)$.*

PROOF. Suppose every reason in $\mathcal{D}$ is finite. By Corollary 18.34, we need only show that $\Lambda_\omega = E$ implies $E \in FGExts(S)$. Suppose $\Lambda_\omega = E$. We first show $\Lambda_\omega = \Lambda$. Suppose, by way of contradiction, that $\Lambda \neq \Lambda_\omega$. Then there must be a least ordinal $\alpha \geq \omega$ such that for some $e \in \mathcal{D}$, $e \in \Lambda_{\alpha+1} - \Lambda_\alpha$. Since α is minimal, $\Lambda_\omega = \Lambda_\alpha$, for otherwise $\Lambda_\omega = \Lambda_{\omega+1}$ and hence $\Lambda_\omega = \Lambda$. By construction, there is some $f \in \Lambda_\alpha$, $\mathcal{I}(f) = A \setminus\setminus B \mid\!\!\!\vdash C$, $A \subseteq \Lambda_\alpha$, $E \subseteq B^c$, and $e \in C$. Since A is finite, this means the rank of A is also finite. Thus there is some $\beta < \omega$ such that $A \subseteq \Lambda_\beta$ and $f \in \Lambda_\beta$, so $e \in \Lambda_{\beta+1} \subseteq \Lambda_\omega$, a contradiction. Thus $\Lambda = \Lambda_\omega$, and since $\Lambda_\omega = E$, by Theorem 18.35 E is a grounded extension of S. We see that E is finitely grounded by induction on rank. Clearly, if $e \in \Lambda_0$, then $e \in S$, hence $\{e\}$ is a rank-preserving grounding set. Now suppose the rank of e is $\alpha + 1 < \omega$. Then by construction there is some $f \in \Lambda_\alpha$ with $\mathcal{I}(f) = A \setminus\setminus B \mid\!\!\!\vdash C$, $A \subseteq \Lambda_\alpha$, $E \subseteq B^c$, and $e \in C$. By inductive hypothesis, each of f and A have finite rank-preserving grounding sets, so merge these preserving rank-order, add e to the end, and the result is a finite rank-preserving grounding order for e. ∎

(18.38) QUESTION. *For each simple reasons agent, can one characterize those sets which have no admissible extensions? Unique admissible extensions? Multiple admissible extensions? Finitely many admissible extensions? Infinitely many admissible extensions? Only finite admissible extensions? Only infinite admissible extensions? Or, turning the order around, can one characterize simple reasons agents in which all sets are coherent or have finite or unique admissible extensions?*

(18.39) THEOREM (GROUNDED MINIMALITY). $GExts \subseteq \mu QExts$.

PROOF. $GExts \subseteq QExts$ by Theorem 18.7, so suppose $E \in GExts(S)$, $E' \in QExts(S)$, and $E' \subseteq E$. We first show $\Lambda(S,E) \subseteq \Lambda(S,E')$ by induction. Clearly $\Lambda_0(S,E) \subseteq \Lambda_0(S,E')$ since each equals S. Assume $\Lambda_\beta(S,E) \subseteq \Lambda_\beta(S,E')$ for each $\beta < \alpha$. If α is a limit ordinal, then by definition $\Lambda_\alpha(S,E') \subseteq \Lambda_\alpha(S,E)$. If α is a successor ordinal, say $\alpha = \beta+1$, let $e \in \Lambda_\alpha(S,E)$. If $e \in S$, then $e \in E'$, and if $e \notin S$ there is a $d \in \Lambda_\beta(S,E)$ with $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\sim C$, $A \subseteq \Lambda_\beta(S,E)$, $E \subseteq B^c$, and $e \in C$. But since $E' \subseteq E \subseteq B^c$, this means $e \in \Lambda_\alpha(S,E')$. Hence $\Lambda(S,E) \subseteq \Lambda(S,E')$. But by Theorem 18.31 and Lemma 18.30, $E = \Lambda(S,E) \subseteq \Lambda(S,E') \subseteq E'$. ∎

(18.40) COROLLARY. *$GExts = \mu GExts$ and $FGExts = \mu FGExts$.*

(18.41) COROLLARY. *If $S \lhd \mathcal{D}$, then $AExts(S) = \{\mathcal{D}\}$.*

PROOF. Suppose $S \lhd \mathcal{D}$. Now if $E \in AExts(S)$, then $E \subseteq \mathcal{D}$, and since $AExts = \mu AExts$ by the previous corollary, this means $E = \mathcal{D}$. ∎

(18.42) COROLLARY. *If $S \lhd \mathcal{D}$, then $A \mid\!\!=_S B$ for every $A, B \subseteq \mathcal{D}$.*

(18.43) COROLLARY. *If $A \mid\!\sim_S B$ for every $A, B \subseteq \mathcal{D}$, then $A \mid\!\!=_S B$ for every $A, B \subseteq \mathcal{D}$.*

Note that the converse does not hold if S is incoherent.

(18.44) THEOREM (TRIVIAL COHERENCE). *A trivially interpreted set is its own admissible extension.*

PROOF. Suppose $S \subseteq \mathcal{D}$, and for all $d \in S$, $\mathcal{I}(d) = \mathbf{P}\mathcal{D}$. Then S is admissible, and it is clearly finitely grounded, so $AExts(S) = \{S\}$. ∎

(18.45) DEFINITION. *A monotonic reason is an element $d \in \mathcal{D}$ such that for some $A, C \subseteq \mathcal{D}$, $\mathcal{I}(d) = A \setminus\setminus \emptyset \mid\!\!\sim C$.*

"Monotonic" is used because extensions of a state cannot invalidate such a reason, so that the set of conclusions added by the reason is monotone nondecreasing with enlargements of the state. Note that all trivially interpreted state components are finite and monotonic.

(18.46) THEOREM (MONOTONIC COHERENCE). *If $\mathcal{D}$ contains only monotonic reasons, every subset of $\mathcal{D}$ has a unique grounded extension.*

PROOF. Suppose all reasons are monotonic, and let $S \subseteq \mathcal{D}$. Consider $E = \Lambda(S,\emptyset)$. Since all reasons are monotonic, $\Lambda(S,\emptyset) = \Lambda(S,X)$ for each $X \subseteq \mathcal{D}$. In particular, $E = \Lambda(S,\emptyset) = \Lambda(S,E)$, so by Theorem 18.35 $E \in GExts(S)$. Now if $E' \in GExts(S)$, then $E' = \Lambda(S,E') = E$ by the previous observation, so $GExts(S) = \{E\}$. ∎

(18.47) COROLLARY. *If $\mathcal{D}$ contains only finite monotonic reasons, every subset of $\mathcal{D}$ has a unique finitely grounded extension.*

As Example 18.24 shows, we cannot drop the finiteness restriction unless we allow infinite grounded arguments.

(18.48) DEFINITION. *Suppose $S \subseteq \mathcal{D}$ and $d \in \mathcal{D}$. The set S mentions d iff for some $e \in S$, $\mathcal{I}(e) = A \setminus\setminus B \mid\!\!\sim C$ and $d \in A \cup B \cup C$. Two sets $A, B \subseteq \mathcal{D}$ have disjoint mention sets iff A mentions no $b \in B$ and B mentions no $a \in A$. A subset $A \subseteq S$ is an isolated subset of S iff A and $S - A$ have disjoint mention sets.*

Note that S and $\emptyset$ are isolated subsets of S.

(18.49) THEOREM (DISJOINT SUM). *Suppose $\langle S_i\rangle_{i=0}^n$ and $\langle \hat{E}_i\rangle_{i=0}^n$ ($n \leq \infty$) are sequences such that for all i, $0 \leq i \leq n$, $S_i \subseteq \mathcal{D}$ and $\hat{E}_i = AExts(S_i)$. Suppose further that for all i, j, if $0 \leq i \neq j \leq n$, then each $E \in \hat{E}_i$ and $E' \in \hat{E}_j$ have disjoint mention sets. Then*

$$AExts(\bigcup_{i=0}^{n} S_i) = \{\bigcup_{k=0}^{n} E^k \mid \langle E^k\rangle_{k=0}^n \in \prod_{i=0}^{n} \hat{E}_i\}.$$

PROOF. Let $E = \bigcup_k E^k$ for some $\langle E^k\rangle \in \Pi_i \hat{E}_i$. We first check the admissibility of E. Let $e \in E$. If e is trivially interpreted, then $E \in \mathcal{I}(e)$, and if e has a nontrivial interpretation, then $e \in E^l$ for some l. Now $E^l \in \mathcal{I}(e)$ by hypothesis, and since the other parts of E are not mentioned by E^l, they also satisfy e, hence E is admissible. Now if $e \notin \bigcup_i S_i$, then it is in some E^l. By hypothesis, e has a finite grounding argument in E^l, and since it does not mention the rest of E, this same argument grounds e in E. Hence E is a finitely grounded extension of S, so $E \in AExts(\bigcup_i S_i)$. Now suppose $E \in AExts(\bigcup_i S_i)$, and consider $E \cap \bigcup \hat{E}_i$. No element in this set mentions any other in $\bigcup \hat{E}_j$ for $j \neq i$, so it must be that $E \cap \bigcup \hat{E}_i \in AExts(S)$. ∎

(18.50) COROLLARY. *$S \subseteq \mathcal{D}$ is incoherent if S has an incoherent isolated subset.*

PROOF. Suppose $A \subseteq S \subseteq \mathcal{D}$, A is isolated in S, and $AExts(A) = \emptyset$. Then $AExts(A) \times AExts(S - A) = \emptyset$, so by the disjoint sum theorem, $AExts(S) = \emptyset$. ∎

(18.51) COROLLARY. *Suppose $S = \bigcup_i S_i$, where for each i, $AExts(S_i) = \{E_i\}$, and if $j \neq i$, E_i and E_j have disjoint mention sets. Then $AExts(S) = \{\bigcup_i E_i\}$.*

PROOF. Since each S_i has a unique extension, there is only one sequence in the product of the extension sets, so the claim follows by the disjoint sum theorem. ∎

(18.52) COROLLARY. *Suppose $S = \bigcup_i S_i$ and if $j \neq i$, E_i and E_j have disjoint mention sets. Then $|AExts(S)| > 1$ if for some i, $|AExts(S_i)| > 1$.*

PROOF. Suppose S_i has several admissible extensions. Then the product of the extension sets contains several sequences. The unions of these sequences cannot be identical, since by the disjointness of mention sets this would mean the supposedly distinct extensions of S_i were identical. Thus by the disjoint sum theorem, the union of the sets has several admissible extensions. ∎

(18.53) COROLLARY. *S is coherent iff every isolated subset of S is coherent.*

(18.54) COROLLARY. *S has a unique extension iff every isolated subset of S has a unique extension.*

(18.55) DEFINITION. *A set $S \subseteq \mathcal{D}$ is called simple iff S has no isolated subsets other than itself and $\emptyset$.*

For example, if $\mathcal{D} = \{a\}$, then $\mathcal{D}$ is simple, and if $\mathcal{D} = \{a, b\}$ where both elements have trivial interpretations, then $\mathcal{D}$ is not simple, since each of $\{a\}$ and $\{b\}$ is.

(18.56) LEMMA. *There are incoherent simple sets.*

PROOF. As before, if $\mathcal{I}(a) = \emptyset \setminus\setminus \{b\} \Vdash \{c\}$, $\mathcal{I}(b) = \mathcal{I}(c) = \emptyset \setminus\setminus \emptyset \Vdash \emptyset$, and $\mathcal{I}(d) = \{c\} \setminus\setminus \emptyset \Vdash \{b\}$, then $\{a, d\}$ is simple but incoherent. ∎

(18.57) THEOREM. *For each $n \in \mathbf{N}$ there is a simple set $S \subseteq \mathcal{D}$ with $|AExts(S)| = n$.*

PROOF. If $n = 0$, the simple set must be incoherent, and the preceding lemma applies. If $n > 0$, let D and E be two disjoint sets with $|D| = |E| = n$, $D = \{d_i\}_{i=1}^n$ and $E = \{e_i\}_{i=1}^n$. For each i, $1 \leq i \leq n$, let $\mathcal{I}(d_i) = \emptyset \setminus\setminus E - \{e_i\} \Vdash \{e_i\}$ and $\mathcal{I}(e_i) = \emptyset \setminus\setminus \emptyset \Vdash \emptyset$. Then each d_i mentions E, so D is simple. But clearly, $AExts(D) = \{D \cup \{e_i\} \mid 1 \leq i \leq n\}$. Hence $|AExts(D)| = n$. ∎

Rather than pursue this structure theory further here, we merely mention the possibility of deriving sufficient conditions on coherence from MCDERMOTT'S termination theorem for RMS.[13] (RMS is an artificial intelligence system for maintaining a database by means of reasons. It is discussed in §46.)

By a *finite agent* we mean one with finite domain $\mathcal{D}$. Necessarily, all reasons of a finite agent are finite as well. Suppose for the time being that $(\mathcal{D},\mathcal{I})$ characterizes a simple reasons agent and $|\mathcal{D}| = n < \omega$. We then have $|\mathcal{I}| = \mathcal{O}(n^2)$, and the following results.

(18.58) THEOREM. *Is $S \in \mathscr{S}$? can be computed in time $\mathcal{O}(n^2)$.*

PROOF. Checking the interpretation of each element of S requires $\mathcal{O}(n)$ steps, and S can have up to n elements. ∎

(18.59) THEOREM. *Is $E \in Exts(S)$? can be computed in time $\mathcal{O}(n^2)$.*

PROOF. Checking $E \supset S$ requires $\mathcal{O}(|E|+|S|) = \mathcal{O}(n)$ steps, and checking admissibility of E requires $\mathcal{O}(n^2)$ steps as above. ∎

(18.60) THEOREM. *Is $E \in QExts(S)$? can be computed in time $\mathcal{O}(n^2)$.*

PROOF. Checking $E \in Exts(S)$ requires $\mathcal{O}(n^2)$ steps, as above. We can check component-admissibility of E at the same time by marking the consequences of each valid reason encountered in checking the admissibility of E, and when done scanning E to see if all unmarked elements are in S. This is also $\mathcal{O}(n^2)$. ∎

(18.61) THEOREM. *Is $E \in FGExts(S)$? can be computed in time $\mathcal{O}(n^3)$.*

PROOF. Consider the following algorithm.

1. $\Lambda_0 \leftarrow S$, $i \leftarrow 0$
2. While $\Lambda_i \neq \Lambda_{i-1}$ do
    3. $i \leftarrow i+1$
    4. $\Lambda_i \leftarrow \Lambda_{i-1}$
    5. For each $e \in \Lambda_{i-1}$ do
        6. $\mathcal{I}(e) = A \setminus\setminus B \mid\!\!- C$
        7. If $A \subseteq \Lambda_{i-1}$ and $E \subseteq B^c$ then $\Lambda_i \leftarrow \Lambda_i \cup C$.
8. Return $E = ?\Lambda_i$.

The algorithm clearly answers $E \in FGExts(S)$, since by Corollary 18.36 $E \in FGExts(S)$ iff $E = \Lambda$, and $\Lambda_i = \Lambda$ whenever $\Lambda_i = \Lambda_{i+1}$. Steps 1, 4, 6, 7 and 8 cost $\mathcal{O}(n)$. The iteration of step 5 may run $\mathcal{O}(n)$ times, so the cost of 5-7 is $\mathcal{O}(n^2)$. Now note that since $|\mathcal{D}| = n$, the loop of step 2 can run at most $n+1$ times, so the total cost is $\mathcal{O}(n) + \mathcal{O}(n) \cdot (\mathcal{O}(n) + \mathcal{O}(n^2)) = \mathcal{O}(n^3)$. ∎

(18.62) THEOREM. *Find $E \in Exts(S)$! is in P, Find $E \in QExts(S)$! and Find $E \in FGExts(S)$! are in NP, and Count $Exts(S)$!, Count $QExts(S)$!, and Count $FGExts(S)$! are in #P.*

PROOF. $\mathcal{D} \in Exts(S)$ for every S (as we observe later in Theorem 20.1), so finding an extension is trivial. The other five problems may all be computed by guessing sets $E \subseteq \mathcal{D}$ and accepting iff the desired condition is true, all deterministic polynomial computations from the above theorems. ∎

(18.63) CONJECTURE. *Find $E \in QExts(S)$! is NP-complete.*

(18.64) QUESTION. *Is Find $E \in FGExts(S)$? NP-complete?*

[13] [CHARNIAK, RIESBECK, AND MCDERMOTT 1980]

(18.65) THEOREM. *Is $E \in \mu Exts(S)$?, Is $E \in \mu QExts(S)$?, and Is $\mu QExts(S) = \emptyset$? are in co-NP.*

PROOF. We see $E \notin \mu Exts(S)$ is in NP by first checking $E \in Exts(S)$ deterministically as above, then guessing a proper subset E' of E, checking if it is in $Exts(S)$ as well, and accepting if either $E \notin Exts(S)$ or $E' \in Exts(S)$. The case of $E \in \mu QExts(S)$ is the same except for testing $E, E' \in QExts(S)$. For the last question, we guess $E \supset S$, check if $E \in QExts(S)$, and accepting if so, since $\mu QExts(S)$ will be nonempty iff $QExts(S)$ is empty. ∎

(18.66) CONJECTURE. *The following problems are NP-hard: Is $E \in \mu Exts(S)$?, Is $E \in \mu QExts(S)$?, Is $\mu QExts(S) = \emptyset$?, and Find $E \in \mu QExts(S)$!*

The development of efficient algorithms for deciding these questions is still the subject of study. The most studied question is that of constructing an admissible extension of a set if one exists. This is one of the tasks of RMS and its relatives. The hard case of course is when no admissible extension exists. The known algorithms typically discover this by exhaustive failure. If a suitably mechanizable characterization of coherence was known, more efficient algorithms might be possible.

(18.67) QUESTION. *Are there interesting classes of simple reasons agents for which construction of admissible extensions is tractable?*

We now drop the supposition that $\mathcal{D}$ is finite and replace it with the notion of a finite "virtual" domain with respect to some subset of $\mathcal{D}$.

(18.68) DEFINITION. *The universe $\mathcal{U}(S)$ of a set S is the smallest set containing S and containing the mention sets of each of its elements, that is, $S \subseteq \mathcal{U}(S)$ and if $d \in \mathcal{U}(S)$ and $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!- C$, then $A, B, C \subseteq \mathcal{U}(S)$.*

Note that if $\mathcal{D}$ is finite, then every $S \subseteq \mathcal{D}$ has a finite universe; that if S has a finite universe, then all reasons in S are finite; and that two sets with disjoint universes have disjoint mention sets, though the converse need not be true.

(18.69) CONJECTURE. *Is $\mathcal{U}(S)$ finite? for finite S is undecidable.*

(18.70) LEMMA. *If $E \in FGExts(S)$, then $E \subseteq \mathcal{U}(S)$.*

PROOF. Let $E \in FGExts(S)$. Then $E = \Lambda_\omega(S, E)$, and we prove $\Lambda_\omega \subseteq \mathcal{U}(S)$ by induction. Since $\Lambda_0 = S$, $\Lambda_0 \subseteq \mathcal{U}(S)$, so suppose $\Lambda_\alpha \subseteq \mathcal{U}(S)$ and consider $\Lambda_{\alpha+1}$. Any $e \in \Lambda_{\alpha+1} - \Lambda_\alpha$ is supported, and hence mentioned by a valid reason in Λ_α, so $\Lambda_{\alpha+1} \subseteq \mathcal{U}(S)$ too. Thus $\Lambda_\omega = E \subseteq \mathcal{U}(S)$. ∎

(18.71) THEOREM. *If S has a finite universe, then admissible extensions of S are decidable.*

PROOF. Compute $\mathcal{U}(S)$, check $E \subseteq \mathcal{U}(S)$, and check $E \in FGExts(S)$ by the earlier methods for finite $\mathcal{D}$. ∎

(18.72) COROLLARY. *If $S \cup \{d\}$ has a finite universe, then both the arguability and inevitability of d in S are decidable.*

(18.73) THEOREM. *If $\mathcal{D}$ is recursively enumerable and the set of nontrivial reasons in $\mathcal{D}$ has a finite universe, then the finite admissible states are recursively enumerable.*

PROOF. Let $\mathcal{U}$ be the universe of the set of all nontrivial reasons. By the previous methods the admissible states $S \subseteq \mathcal{U}$ are enumerable. Also, all finite subsets of $\mathcal{D} - \mathcal{U}$ are enumerable, so we can by composition enumerate the sets $S \cup U$, where $S \subseteq \mathcal{U}$, $S \in \mathscr{S}$, and $U \subseteq \mathcal{D} - \mathcal{U}$, and so exhaust the finite elements of $\mathscr{S}$. ∎

§**19.** Our second theory of reasoned assumptions modifies the simple reasons theory by interpreting state components as *defeasible reasons*. To do this, we suppose a function *Defeated* : $\mathcal{D} \to \mathcal{D}$ and interpret the element *Defeated*(d) to mean the reason d has been defeated and cannot support conclusions. This is achieved by trivializing the interpretation of d in the presence of *Defeated*(d), specifically, for each $d \in \mathcal{D}$ having

$$\mathcal{I}(d) \supset \{S \subseteq \mathcal{D} \mid \textit{Defeated}(d) \in S\}.$$

In effect, we interpret d in different ways depending on whether the state containing it also contains *Defeated*(d) or not.

(19.1) DEFINITION. *Defeasible reasons agents are characterized by the following additions to the requirements on simple reasons agents:*

(i) There is a 1-1 function Defeated : $\mathcal{D} \to \mathcal{D}$,

(ii) For each $d \in \mathcal{D}$ *if* $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\sim C$, *then Defeated*$(d) \in B$.

Thus if $\mathcal{I}(d) = A \setminus\setminus B^+ \mid\!\!\sim C$, where $B^+ = B \cup \{\textit{Defeated}(d)\}$, we have by simple rewriting

$$\begin{aligned} A \subseteq S \subseteq (B^+)^c \supset C \subseteq S \quad &\text{iff} \quad [A \subseteq S \subseteq B^c \wedge \textit{Defeated}(d) \notin S] \supset C \subseteq S \\ &\text{iff} \quad \textit{Defeated}(d) \notin S \supset [A \subseteq S \subseteq B^c \supset C \subseteq S], \end{aligned}$$

which is the desired condition of triviality.

EXAMPLES. We adopt the convention of eliding most mentions of *Defeated* elements in reasons by use of a superscript $^+$ to refer to the appropriate element. We omit for brevity most previous examples, since they are basically unchanged by the introduction of defeasible reasons.

(19.2) $\mathcal{I}^*(S) = \{\emptyset \setminus\setminus \emptyset^+ \mid\!\!\sim A\}, \qquad AExts(S) = \{S \cup A\}$ (as before)

(19.3) $\mathcal{I}^*(S) = \{\emptyset \setminus\setminus \neg A^+ \mid\!\!\sim A\}, \qquad AExts(S) = \{S \cup A\}$ (as before)

(19.4) $\mathcal{I}^*(S) = \{\emptyset \setminus\setminus \emptyset^+ \mid\!\!\sim A, \quad \emptyset \setminus\setminus \emptyset^+ \mid\!\!\sim \neg A\}, \qquad AExts(S) = \{S \cup A \cup \neg A\}$ (as before)

(19.5) $S = \{a, b\}, \quad \mathcal{I}(a) = \emptyset \setminus\setminus \emptyset^+ \mid\!\!\sim A, \quad \mathcal{I}(b) = \emptyset \setminus\setminus \emptyset^+ \mid\!\!\sim \{\textit{Defeated}(a)\},$
$AExts(S) = \{S \cup \{\textit{Defeated}(a)\}\}$

Note how the challenge defeats the otherwise valid reason.

(19.6) $S = \{a, b, c\}, \quad \mathcal{I}(a) = \emptyset \setminus\setminus \emptyset^+ \mid\!\!\sim A, \quad \mathcal{I}(b) = \emptyset \setminus\setminus \emptyset^+ \mid\!\!\sim \{\textit{Defeated}(a)\},$
$\mathcal{I}(c) = \emptyset \setminus\setminus \emptyset^+ \mid\!\!\sim \{\textit{Defeated}(b)\}$
$AExts(S) = \{S \cup A \cup \{\textit{Defeated}(b)\}\}$

Here the first challenge is defeated by the second, so the initial reason's conclusions are unaffected.

As one might suspect, the innocuous reinterpretation of reasons lets us extend any simple reasons theory to a defeasible reasons theory without important changes of meaning.

(19.7) THEOREM (DEFEASIBLE EMBEDDING). *Suppose that* $\mathcal{D}$ *and* $\mathcal{I}$ *characterize a simple reasons agent. Then there is a defeasible reasons agent* $(\mathcal{D}', \mathcal{I}')$ *such that* $\mathcal{D} \subseteq \mathcal{D}'$ *and if* $S \subseteq \mathcal{D}$, *the admissible extensions of* S *in* $(\mathcal{D}, \mathcal{I})$ *are exactly the admissible extensions of* S *in* $(\mathcal{D}', \mathcal{I}')$.

PROOF. Let $(\mathcal{D}, \mathcal{I})$ describe a simple reasons agent. We define the desired defeasible reasons agent $(\mathcal{D}', \mathcal{I}')$ as follows. Let $\mathcal{D}' = \mathcal{D}_1 \oplus \mathcal{D}_2$, where $\mathcal{D} = \mathcal{D}_1 = \mathcal{D}_2$. We always say $d \in \mathcal{D}$ to mean $d \in \mathcal{D}_1$, and write the corresponding element of $\mathcal{D}_2$ as d'. If $d \in \mathcal{D}$ and $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\sim C$, define $\mathcal{I}'(d) = A \setminus\setminus B \cup \{d'\} \mid\!\!\sim C$. If $d' \in \mathcal{D}'$, define $\mathcal{I}'(d') = \emptyset \setminus\setminus \{d\} \mid\!\!\sim \emptyset$ (the trivial interpretation). Now suppose $S \subseteq \mathcal{D}$ and $E \subseteq \mathcal{D}'$. We claim $E \in AExts(S)$ iff $E \in AExts'(S)$. First, suppose $E \in AExts(S)$. Then $E \subseteq \mathcal{D}$ and $E \cap \mathcal{D}_2 = \emptyset$, so for each $d \in E$, $E \in \mathcal{I}'(d)$, hence $E \in \mathscr{S}$. But clearly all grounding arguments in $\mathcal{D}$ then translate directly into grounding arguments in $\mathcal{D}'$, so $E \in AExts'(S)$. Second, suppose instead that $E \in AExts'(S)$. Clearly d' cannot occur in E unless $d' \in S$, which is not the case, so $E \cap \mathcal{D}_2 = \emptyset$. Because of this, $\mathcal{I}'(d) = \mathcal{I}(d)$ for each $d \in E$, so $E \in \mathscr{S}$, and every grounding argument in E translates directly into a grounding argument in $\mathcal{D}$, so $E \in AExts(S)$. ∎

(19.8) THEOREM (MONOTONE EMBEDDING). *Suppose* $(\mathcal{D}, \mathcal{I})$ *characterizes a simple reasons agent, and that* $\langle S_i \rangle_{i=0}^n$ $(n \leq \omega)$ *is a sequence of subsets of* $\mathcal{D}$. *Then there is a defeasible reasons agent* $(\mathcal{D}', \mathcal{I}')$ *and a sequence*

$\langle S'_i\rangle_{i=0}^n$ of monotone nondecreasing subsets of $\mathcal{D}'$ such that for each i there is a 1-1 correspondence f between the admissible extensions of S_i and S'_i such that $E = f(E) \cap \mathcal{D}$ whenever $S_i \lhd E$.

PROOF. Suppose $(\mathcal{D}, \mathcal{I})$ characterizes a simple reasons agent, and that $\langle S_i\rangle_{i=0}^n$ ($n \leq \omega$) is a sequence of subsets of $\mathcal{D}$. We define the desired defeasible reasons agent $(\mathcal{D}', \mathcal{I}')$ as follows. Let $\mathcal{D}' = \mathcal{D} \oplus (\mathcal{D} \times (\omega + 1))$. We use the elements (d, n) for finite n to encode the addition or removal of the element d as follows.

For each $d \in \mathcal{D}$, $Defeated(d) = (d, \omega)$, $Defeated((d, \omega)) = d$, and for each $k \in \mathbf{N}$, $Defeated((d, 2k)) = (d, 2k + 1)$, and $Defeated((d, 2k + 1)) = (d, 2k)$.

For each $d \in \mathcal{D}$, if $\mathcal{I}(d) = A \setminus\setminus B \mid\!\sim C$, then $\mathcal{I}'(d) = A \setminus\setminus B \cup \{(d, \omega)\} \mid\!\sim C$. Also, $\mathcal{I}'((d, \omega)) = \emptyset \setminus\setminus \{d\} \mid\!\sim \emptyset$ (the trivial interpretation), $\mathcal{I}'((d, 0)) = \emptyset \setminus\setminus \{(d, 1)\} \mid\!\sim \{d\}$, and for each $k \in \mathbf{N}$, $\mathcal{I}'((d, 2k + 1)) = \emptyset \setminus\setminus \{(d, 2k)\} \mid\!\sim \emptyset$ and $\mathcal{I}'((d, 2k + 2)) = \emptyset \setminus\setminus \{(d, 2k + 3)\} \mid\!\sim \{(d, 2k + 1)\}$.

Let $S'_0 = S_0 \times \{0\}$, and if $0 < i \leq n$, let $S'_{i+1} = S'_i \cup \{(d, 2k + 2) \mid d \in S_i \mathbin{\triangle} S_{i+1} \ \wedge \ k = \max\{j \mid (d, 2j) \in S'_i\}\}$. Obviously, the sequence $\langle S'_i\rangle_{i=0}^n$ is nondecreasing, and for each i, $S'_i \cap \mathcal{D} = \emptyset$. We claim for each i, $0 \leq i \leq n$, $|AExts(S_i)| = |AExts'(S'_i)|$. Clearly, no elements of $\mathcal{D} \times \{\omega\}$ can ever occur in admissible extensions of S'_i since no reasons have them as conclusions. Now the element $(d, 0)$ will be defeated iff the last change with respect to d was its removal, so if S_i contains d, $(d, 0)$ will be undefeated in every extension of S'_i, and so include d. These elements then reproduce the extensions of the original set. Since the elements of $\mathcal{D}$ mention no elements of $\mathcal{D} \times (\omega + 1)$, if $E \in AExts'(S'_i)$, then $E \cap \mathcal{D} \in AExts(S_i)$. ∎

§**20.** While the ideas of simple and defeasible reasons play important roles in the practice of artificial intelligence, they have important limitations in the sorts of state specifications they can express. The most obvious of these is the inability to exclude any state component from admissible states.

(20.1) THEOREM. *$\mathcal{D}$ is admissible in the simple reasons theory.*

PROOF. For all $d \in \mathcal{D}$, there are sets $A, B, C \subseteq \mathcal{D}$ such that $\mathcal{I}(d) = A \setminus\setminus B \mid\!\sim C$, and for all sets $A, B, C \subseteq \mathcal{D}$, $C \subseteq \mathcal{D}$, and hence $A \subseteq \mathcal{D} \subseteq B^c \supset C \subseteq \mathcal{D}$, is always true. ∎

Noting this obvious inexpressiveness, it is natural to ask exactly which state spaces $\mathscr{S}$ are expressible in the simple reasons theory. As we noted before, the empty set cannot be excluded.

(20.2) THEOREM. *$\emptyset$ is admissible in the simple reasons theory.*

Beyond such trivialities, the expressive capabilities of the simple reasons theory are largely unknown. Yet if we return to the unrestricted theory, we have clearer expressive powers.

(20.3) THEOREM. *Let $\mathscr{S} \subseteq \mathbf{P}\,\mathcal{D}$. Then $\mathcal{I}$ can be chosen so that $\mathcal{Q} = \mathscr{S} \cup \{\emptyset\}$.*

PROOF. Define $\mathcal{I}$ so that for each $d \in \mathcal{D}$, $\mathcal{I}(d) = \mathscr{S}$. ∎

Since the general theory has unlimited but practically unusable expressive capabilities, while the simple reasons theory is somewhat inexpressive in spite of its demonstrated practical utility, we look for additions to the simple reasons theory which increase the expressive capabilities in useful ways. The most obvious candidate addition is the notion of *denial*.

(20.4) DEFINITION. *An agent's use of simple reasons and denials is characterized exactly as in the simple reasons theory, except that $\mathcal{D} = \mathcal{D}^+ \oplus \mathcal{D}^-$, where for each $d \in \mathcal{D}^+$ there are sets $A, B, C \subseteq \mathcal{D}$ such that*

$$\mathcal{I}(d) = A \setminus\setminus B \mid\!\sim C = \{S \subseteq \mathcal{D} \mid A \subseteq S \subseteq B^c \supset C \subseteq S\},$$

and for each $d \in \mathcal{D}^-$ there is some (necessarily unique) $e \in \mathcal{D}$ such that

$$\mathcal{I}(d) = \{S \subseteq \mathcal{D} \mid e \notin S\}.$$

Note that this theory is just like the simple reasons theory except that some state components, the denials, can rule out the presence of specific elements. To make this ability uniform, we need only assume the existence of a 1-1 function $\neg : \mathcal{D} \to \mathcal{D}^-$ such that for each $d \in \mathcal{D}$, $\mathcal{I}(\neg d) = \{S \subseteq \mathcal{D} \mid d \notin S\}$. This augmentation of the theory adds expressive power, but again we have no characterization of exactly which state spaces are so expressible.

(20.5) THEOREM. *If $\mathcal{D}$ is infinite, then not every state space can be attained with only finite reasons and denials.*

PROOF. If $|\mathcal{D}| = \omega$, then the number of possible finite reasons is ω, as is the number of possible denials. Hence the number of possible interpretation functions is $(\omega + \omega)^\omega = \omega^\omega$. But the number of possible state spaces is 2^{2^ω}, and $\omega^\omega < 2^{2^\omega}$. ∎

One should not confuse the nature or uses of denials with the qualifiers of simple reasons or the defeaters of defeasible reasons. Their natures differ in that the former sort of element absolutely rules out some conclusion from appearing in any admissible extension, while the latter sorts of elements permit the reasons's conclusions to appear if supported by other reasons. The uses of these notions differ as well. Inclusion of a denial in a set rules out some element with no recourse, and for "no reason" other than the denial itself. In contrast, the other notions permit reasoned retraction of assumptions, defeating an assumption for a reason, and possibly later restoring the assumption if the defeating reason itself is defeated.

One special case to consider is when the agent admits *self-denials* or *contradictions*. Such are elements denying themselves, and are necessarily inadmissible, since $\mathcal{I}(d) = \{S \subseteq \mathcal{D} \mid d \notin S\}$. Such elements are useful in practice, as we see later.

(20.6) THEOREM. *Every agent using simple reasons and denials can be realized within an agent using only simple reasons and a single self-denial.*

PROOF. Suppose $\mathcal{D} = \mathcal{D}^+ \oplus \mathcal{D}^-$ and $\mathcal{I}$ characterize an agent's use of simple reasons and denials. Let $\perp$ be some object not in $\mathcal{D}$. Let $\mathcal{D}' = \mathcal{D} \cup \{\perp\}$. Define $\mathcal{I}'$ for each $d \in \mathcal{D}'$ by

(1) $\mathcal{I}'(\perp) = \emptyset$,
(2) If $d \in \mathcal{D}^+$, then $\mathcal{I}'(d) = \mathcal{I}(d)$,
(3) If $d \in \mathcal{D}^-$ and $\mathcal{I}(d) = \{S \subseteq \mathcal{D} \mid e \notin S\}$, then $\mathcal{I}'(d) = \{e\} \setminus\setminus \emptyset \mid\!\!\sim \{\perp\}$.

Let $\mathscr{S}$ and $\mathscr{S}'$ be the respective sets of admissible states. We claim $\mathscr{S} = \mathscr{S}'$. Clearly, $\perp \notin S$ for each $S \in \mathscr{S}'$, so both $\mathscr{S}$ and $\mathscr{S}'$ are subsets of $\mathbf{P}\,\mathcal{D}$. Suppose $d \in \mathcal{D}^+$. Then by definition, $\mathcal{I}(d) \subseteq \mathcal{I}'(d)$. Suppose $d \in \mathcal{D}^-$ is a denial of $e \in \mathcal{D}$. Then $\mathcal{I}'(d) = \{e\} \setminus\setminus \emptyset \mid\!\!\sim \{\perp\} = \{S \subseteq \mathcal{D} \mid e \in S \supset \perp \in S\} = \mathcal{I}(d) \cup \{S \subseteq \mathcal{D}' \mid \perp \in S\} \supset \mathcal{I}(d)$. Thus if $d \in \mathcal{D}$, $\mathcal{I}(d) \subseteq \mathcal{I}'(d)$, and if $S \subseteq \mathcal{D}$, $\bigcap_{d \in S} \mathcal{I}(d) \subseteq \bigcap_{d \in S} \mathcal{I}'(d)$, so $S \in \mathscr{S}$ only if $S \in \mathscr{S}'$. Suppose $S \in \mathscr{S}'$. Then $\perp \notin S$, so if $S \in \mathcal{I}'(d)$, $S \in \mathcal{I}(d)$, hence $S \in \mathscr{S}$. ∎

(20.7) THEOREM. *Let $\mathscr{S} \subseteq \mathbf{P}\,\mathcal{D}$, $\perp \notin \mathcal{D}$, and $\mathcal{D}' = \mathcal{D} \cup \{\perp\}$. Then $\mathscr{S} \cup \{\emptyset\}$ is realizable as the set of admissible states employing simple reasons $\mathcal{D}$ and contradiction $\perp$ if a distinct representative can be chosen from each set in $\mathbf{P}\,\mathcal{D} - \mathscr{S} - \{\emptyset\}$.*

PROOF. If $S \in \mathbf{P}\,\mathcal{D} - \mathscr{S} - \{\emptyset\}$, let $R(S) \in S$ be the selected distinct representative. Define $\mathcal{I}$ so that $\mathcal{I}(d) = \mathbf{P}\,\mathcal{D}$ if d is not one of the representatives, and $\mathcal{I}(d) = S \setminus\setminus \mathcal{D} - S \mid\!\!\sim \{\perp\}$ if $d = R(S)$. We claim $S \subseteq \mathcal{D}$ is admissible iff $S \in \mathscr{S} \cup \{\emptyset\}$. Now $\mathcal{I}(R(S)) = \mathbf{P}\,\mathcal{D}' - \{S\}$, since $S \subseteq X \subseteq (\mathcal{D} - S)^c \supset \perp \in X$ is false only when $X = S$. Thus S is inadmissible, since $R(S) \in S \notin \mathcal{I}(R(S))$, so S is inadmissible whenever $S \notin \mathscr{S} \cup \{\emptyset\}$. Now $\emptyset$ is admissible as usual, so suppose $d \in S \in \mathscr{S}$. If d is not a selected representative of any set, $S \in \mathcal{I}(d)$ by definition. But if $d = R(S')$, then $S \neq S'$, so $S \in \mathcal{I}(d)$. Hence in either case, $S \in \mathcal{I}(d)$, so S is admissible. ∎

Unfortunately, the complete expressivity of the theory of simple reasons with a contradiction is useless in practice, since the system of distinct representative of the nontrivial inadmissible states is little more than a list of all the inadmissible states, and that is about as useful as the list of all admissible states used in the interpretations of Theorem 20.3.

(20.8) THEOREM. *If d is a denial, but not a self-denial, then $\{d\}$ is admissible.*

PROOF. Suppose $\mathcal{I}(d) = \{S \subseteq \mathcal{D} \mid e \notin S\}$ and $e \neq d$. Then $\{d\} \in \mathcal{I}(d)$, so $\{d\}$ is admissible. ∎

Note that if d is a denial of e, and e a denial of d, then any element c such that $\mathcal{I}(c) = \emptyset \setminus\setminus \emptyset \mid\!\!\vdash C$ with $d, e \in C$ is effectively a self-denial.

§**21.** The theories of simple and defeasible reasons capture the essence of reasoned assumptions, yet are quite general since little restriction was placed on the nature of possible state components. Indeed, though it would divert us too much here, one can develop these theories as theories of reasons invariant under arbitrary bijections of domains. Since the structure of a logical language need not be preserved under arbitrary automorphisms, we see that the general theories are more widely applicable than ones formulated in terms of logical languages. But logical languages play an important role in artificial intelligence, since most mechanized agents constructed to date are explicitly representational, so we continue the development of our theories by examining the special case of representational domains. By representational we mean domains isomorphic with logical languages, that is, systems invariant under all bijections which preserve the logical structure of state components. Since we allow arbitrary re-representation, we may choose a particular logical language as our *reference language* for the purpose of setting out the special logical theories of reasoned assumptions. This we do shortly, but first we treat an important non-linguistic notion that also enters into logical systems.

§**22.** The theory of reasoned assumptions with *deductively closed states* adds to the simple reasons theory the constitutive assumption that all states be closed with respect to a "deducibility" relation and that admissible extensions allow "deductions" in grounding arguments as well as valid reasons. The notion of deducibility we employ is more abstract than the familiar one of deducibility in logic.

(22.1) DEFINITION. *A deducibility relation in $\mathcal{D}$ is a relation $\vdash$ on $\mathbf{P}\,\mathcal{D} \times \mathbf{P}\,\mathcal{D}$ satisfying*

(i) $A \vdash A$ *(reflexivity)*
(ii) If $A \subseteq B$ *and* $A \vdash C$*, then* $B \vdash C$ *(monotonicity)*
(iii) If $A \vdash B$ *and* $A \vdash C$*, then* $A \vdash B \cup C$ *(additivity)*
(iv) If $A \vdash B$ *and* $B \vdash C$*, then* $A \vdash C$ *(transitivity)*
(v) If $A \vdash \{e\}$*, then* $C \vdash \{e\}$ *for some finite* $C \subseteq A$*.* *(compactness)*

We say $S \subseteq \mathcal{D}$ is deductively closed if $B \subseteq S$ whenever $A \subseteq S$ and $A \vdash B$.

Agents whose admissible states must be deductively closed can be thought of as agents which regularly perform some inferences automatically in addition to the inferences involved in making reasoned assumptions, whose problematic character sometimes demands less automatic, more deliberate consideration. By defining deductive closure at this level of abstraction, we allow a wide range of deducibility relations which we can use to characterize a variety of agents with limited automatic inferencing powers. For example, if we take $\vdash$ to be $\supset$, so that $A \vdash B$ iff $A \supset B$, then all sets are "deductively closed," and we have again the simple reasons theory. This corresponds to an agent with no (automatic) inferential resources at all. Even without logical structure for $\mathcal{D}$, the theory of data-types can be cast in terms of deducibility relations.[14] In these theories, sets of components represent partial data-structures, and deductive closure amounts to filling in missing but implied "fields" to further complete the data-structure. If $\mathcal{D}$ has the structure of an ordinary logical language, we might define $A \vdash B$ to hold whenever for each $b \in B$, either $b \in A$ or b is a ground instance of some $a \in A$. In this case, the deductive closure of a set of wffs is just the wffs plus all their ground instances. This corresponds to agents who can automatically instantiate formulas but not combine them. One can go on like this in many ways, for instance capturing agents who can automatically apply Modus Ponens but not instantiate schema or universal statements.

(22.2) THEOREM. *$\mathcal{D}$ is deductively closed.*

PROOF. Suppose $A \subseteq \mathcal{D}$ and $A \vdash B$. Then by definition $B \subseteq \mathcal{D}$, so $\mathcal{D}$ is deductively closed. ∎

[14] [SCOTT 1982]

(22.3) LEMMA. *If every $S \in \hat{S}$ is deductively closed, then $\bigcap \hat{S}$ is deductively closed.*

PROOF. Suppose $A \subseteq \bigcap \hat{S}$ and $A \vdash B$. Then for every $S \in \hat{S}$, $A \subseteq S$, and since S is deductively closed, $B \subseteq S$. But then $B \subseteq \bigcap \hat{S}$. ∎

(22.4) DEFINITION. $\mathrm{Th}(S)$, *the deductive closure of $S \subseteq \mathcal{D}$, is given by*

$$\mathrm{Th}(S) = \bigcap \{S' \mid S \subseteq S' \text{ and } S' \text{ deductively closed}\}.$$

The elements of $\mathrm{Th}(\emptyset)$, *if any, are called tautologies.*

(22.5) COROLLARY. $\mathrm{Th}(S)$ *is the smallest deductively closed superset of S.*

(22.6) THEOREM. *If $A \subseteq B \subseteq \mathcal{D}$, then*
(i) $\mathrm{Th}(A) \subseteq \mathrm{Th}(B)$, *and*
(ii) $\mathrm{Th}(\mathrm{Th}(A)) = \mathrm{Th}(A)$, *and*
(iii) A is deductively closed iff $A = \mathrm{Th}(A)$.

PROOF. (i) Since $A \subseteq B \subseteq \mathrm{Th}(B)$, $\mathrm{Th}(B)$ is a deductively closed superset of A, so by definition, $\mathrm{Th}(A) \subseteq \mathrm{Th}(B)$.
(ii) Since every set is its own superset, and since $\mathrm{Th}(A)$ is deductively closed, by definition we have $\mathrm{Th}(\mathrm{Th}(A)) \subseteq \mathrm{Th}(A)$. But also by definition we have $\mathrm{Th}(A) \subseteq \mathrm{Th}(\mathrm{Th}(A))$, so $\mathrm{Th}(A) = \mathrm{Th}(\mathrm{Th}(A))$.
(iii) *(if)* Immediate since $\mathrm{Th}(A)$ is deductively closed. *(only if)* Since A is its own superset, $\mathrm{Th}(A) \subseteq A$. But $A \subseteq \mathrm{Th}(A)$, so $A = \mathrm{Th}(A)$. ∎

We also introduce interdeducibility equivalence classes into $\mathcal{D}$ by defining $[\,] : \mathcal{D} \to \mathbf{P}\,\mathcal{D}$ for all $d \in \mathcal{D}$ so that

$$[d] = \{e \in \mathcal{D} \mid \{d\} \vdash \{e\} \ \wedge \ \{e\} \vdash \{d\}\}.$$

Here we write $[d]$ instead of $[\,](d)$. When $\mathcal{D}$ has logical structure, the ranges of $[\,]$ are called Lindenbaum algebras.

(22.7) DEFINITION. *E is a (finitely) grounded extension of S iff $S \lhd E$ and for each $e \in E$ there is a (finite) grounding set $G \subseteq E$ and a well-ordering $<_G$ of G such that $e \in G$ and whenever $d \in G$, either (1) $d \in S$ or (2) there is some (finite) reason $f \in G$ and sets $A, B, C \subseteq \mathcal{D}$ such that $\mathcal{I}(f) = A \setminus\setminus B \mid\!\sim C$, $A \subseteq G$, $E \subseteq B^c$, $d \in C$, and $A <_G f <_G d$, or (3) there is some (finite) set $A \subseteq G$ such that $A <_G d$ and $A \vdash \{d\}$. For each $S \subseteq \mathcal{D}$, $GExts(S)$ is the set of grounded extensions of S, and $FGExts(S)$ is the set of finitely grounded extensions of S.*

Note that this definition of groundedness differs from the earlier definition only by allowing grounding sets to include deductive arguments as well as "axioms" and valid reasons. Indeed, the definition is exactly the same as before if we choose $\vdash$ to mean $\supset$, so this notion extends the earlier one.

(22.8) DEFINITION. *An agent's use of finitely grounded simple reasons in deductively closed states is characterized by the axioms for simple reasons agents with the following modifications:*
(i) $\vdash$ is a deducibility relation in $\mathcal{D}$
(ii) $\mathcal{R} = \{S \subseteq \mathcal{D} \mid S = \mathrm{Th}(S)\}$
(iii) For each $d \in \mathcal{D}$ and $S \subseteq \mathcal{D}$,

$$\mathcal{J}(d, S) = \{E \quad | \quad d \in S \ \vee \ E - \{d\} \vdash \{d\} \ \vee \ \exists e \in E \ \exists A, B, C \subseteq \mathcal{D} \\ \mathcal{I}(e) = A \setminus\setminus B \mid\!\sim C \ \wedge \ d \in C \ \wedge \ A \subseteq E \subseteq B^c\}.$$

Again, this definition extends that of simple reasons agents, for if $\vdash$ is $\supset$, then $\mathcal{R} = \mathbf{P}\,\mathcal{D}$ and $E - \{d\} \nvdash \{d\}$, reducing $\mathcal{J}$ to the earlier definition.

(22.9) THEOREM. *$GExts \subseteq QExts$.*

PROOF. Suppose $E \in GExts(S)$ and $e \in E$ has grounding set G. Then by the definition of finite grounding sets, either $e \in S$, or there is an $A \subseteq G$ such that $e \notin A$ and $A \vdash \{e\}$, hence $E - \{e\} \vdash \{e\}$, or some valid reason in E supports e. ∎

(22.10) DEFINITION. *Let $S, E \subseteq \mathcal{D}$. Then $\langle \Lambda_\alpha \rangle$ (α an ordinal), the levels from S in E, are defined for all ordinals by*

$$\Lambda_0(S, E) = S,$$

$$\Lambda_{\alpha+1}(S, E) = \mathrm{Th}(\Lambda_\alpha) \cup \bigcup \{C \subseteq \mathcal{D} \quad | \quad \exists e \in \Lambda_\alpha(S, E)\ \exists A, B \subseteq \mathcal{D}$$
$$\mathcal{I}(e) = A \setminus\setminus B \mid\!\!\sim C \wedge A \subseteq \Lambda_\alpha(S, E) \wedge E \subseteq B^c\},$$

and for limit ordinals λ,

$$\Lambda_\lambda(S, E) = \bigcup_{\alpha < \lambda} \Lambda_\alpha(S, E).$$

We define $\Lambda(S, E) = \bigcup_\alpha \Lambda_\alpha(S, E)$ to be the sum of all levels.

Note how the levels of a deductively closed agent include the deductive closures of the preceding levels of inference via valid reasons. Not surprisingly, we can extend the previous results for simple reasons agents to the deductively closed case. The reader is invited to skip to Corollary 22.29 while we perform this chore.

(22.11) COROLLARY. *If $\alpha \leq \beta$, then $S \subseteq \Lambda_\alpha(S, E) \subseteq \Lambda_\beta(S, E) \subseteq \Lambda(S, E)$.*

(22.12) COROLLARY. *If $\Lambda_\alpha(S, E) = \Lambda_{\alpha+1}(S, E)$, then $\Lambda_\alpha(S, E) = \Lambda(S, E)$.*

(22.13) THEOREM. *If α is the cardinal number $|\mathcal{D}| + 1$, then $\Lambda(S, E) = \Lambda_\alpha(S, E)$.*

PROOF. Since $\mathcal{D}$ has fewer than α elements, it must be that for some $\beta + 1 \leq \alpha$ no new element is introduced in $\Lambda_{\beta+1}$, in other words, $\Lambda_\beta = \Lambda_{\beta+1}$. But then $\Lambda_\beta = \Lambda$, and since $\Lambda_\beta \subseteq \Lambda_\alpha \subseteq \Lambda$, we have $\Lambda = \Lambda_\alpha$. ∎

(22.14) DEFINITION. *If $e \in \Lambda(S, E)$, the rank of e in $\Lambda(S, E)$ is the least ordinal α such that $e \in \Lambda_\alpha(S, E)$. If $A \subseteq \Lambda(S, E)$, the rank of A in $\Lambda(S, E)$ is the least ordinal not less than the rank of any element of A.*

(22.15) LEMMA. *For every $S, E \subseteq \mathcal{D}$, $\Lambda(S, E) = \mathrm{Th}(\Lambda(S, E))$.*

PROOF. Suppose $A \subseteq \Lambda$. A has rank, say α, so if $A \vdash B$, then $B \subseteq \Lambda_{\alpha+1} \subseteq \Lambda$. ∎

(22.16) LEMMA. *If $S \lhd E$, then $\Lambda(S, E) \subseteq E$.*

PROOF. Let $S \lhd E$. Clearly $\Lambda_0 \subseteq E$, so assume $\Lambda_\beta \subseteq E$ for each $\beta < \alpha$. If α is a limit ordinal, then by definition $\Lambda_\alpha \subseteq E$. If α is a successor ordinal, say $\alpha = \beta + 1$, let $e \in \Lambda_\alpha$. If $e \in S$, then $e \in E$, so suppose $e \notin S$. If $e \in \mathrm{Th}(\Lambda_\beta)$, then $e \in E$ since E is deductively closed. If $e \notin \mathrm{Th}(\Lambda_\beta)$ there is a $d \in \Lambda_\beta$ with $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\sim C$, $A \subseteq \Lambda_\beta$, $E \subseteq B^c$, and $e \in C$. Since E is admissible, this means $C \subseteq E$, so $e \in E$. Hence $\Lambda_\alpha \subseteq E$, so $\Lambda \subseteq E$. ∎

(22.17) THEOREM (STRATIFICATION). *If $E \in GExts(S)$ then $\Lambda(S, E) = E$.*

PROOF. Since $S \lhd E$, by the preceding lemma we have $\Lambda \subseteq E$. To see that $E \subseteq \Lambda$, suppose $e \in E$. Since E is a grounded extension of S, there is a grounding set $G \subseteq E$ for e from S in E. We show $G \subseteq \Lambda$ by $<_G$-induction. Let $f \in G$ have no predecessors in $<_G$. Clearly f is the minimum of G, and by definition of G, we must have $f \in S$, hence $f \in \Lambda$. Now suppose that $f \in G$ and for each $d <_G f$, either $d \in S$ or there is a grounding subargument $G' \subseteq G$ for d. If $f \in S$, then $f \in \Lambda$. If $\{g \in G \mid g <_G f\} \vdash \{f\}$, then $f \in \Lambda$ since Λ is deductively closed. Otherwise there is a $d \in G$ such that $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\sim C$, $A <_G d <_G f$, $E \subseteq B^c$, and $f \in C$. By the inductive hypothesis, $A \subseteq \Lambda$ and $d \in \Lambda$, so there is some ordinal α such that $A \subseteq \Lambda_\alpha$ and $d \in \Lambda_\alpha$. But then by construction $C \subseteq \Lambda_{\alpha+1}$, so $f \in \Lambda$. Hence $E \subseteq \Lambda$, so $E = \Lambda$. ∎

(22.18) COROLLARY. *If $E \in GExts(S)$ and $\alpha = |E| + 1$, then $E = \Lambda_\alpha(S, E)$.*

(22.19) COROLLARY. *If $E \in GExts(S)$ and α is its rank, then $E = \Lambda_\alpha(S, E)$.*

(22.20) COROLLARY. *If $E \in FGExts(S)$, then $E = \Lambda_\omega(S, E)$.*

(22.21) THEOREM (FIXED POINT). *If $E = \Lambda(S, E)$, then $E \in GExts(S)$.*

PROOF. Suppose $\Lambda = E$. Since $S \subseteq \Lambda$, $S \subseteq E$. Let $e \in E$ with $\mathcal{I}(e) = A \setminus\setminus B \mid\!\!\sim C$, and suppose $A \subseteq E$. Then there is an ordinal α such that $e \in \Lambda_\alpha$ and $A \subseteq \Lambda_\alpha$, so by construction if $E \subseteq B^c$ as well, then $C \subseteq \Lambda_{\alpha+1} \subseteq E$. Similarly, if $A \vdash B$, $A \subseteq \Lambda$, and A has rank α, then $B \subseteq \Lambda_{\alpha+1} \subseteq \Lambda$. Thus E is admissible. We prove E is a grounded extension of S by induction on rank. Specifically, we prove that each element of E has a rank-preserving grounding set, a set $G \subseteq E$ such that $\mathit{rank}(a) \leq \mathit{rank}(b)$ whenever $a \leq_G b$. Let $e \in E$ have rank α. If $\alpha = 0$, then $e \in S$ and we are done since $\{e\}$ is a rank-preserving grounding argument for e from S in E. Now assume that $\alpha > 0$ and all elements of rank less than α have rank-preserving grounding arguments. Necessarily, α is a successor ordinal, since no elements are introduced at limit ordinals, so suppose $\alpha = \beta + 1$. If $e \in \mathrm{Th}(\Lambda_\beta)$, then there is some $G \subseteq \Lambda_\beta$ such that $G \vdash \{e\}$, and otherwise there is some $d \in \Lambda_\beta$ such that $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\sim C$, $A \subseteq \Lambda_\beta$, $E \subseteq B^c$, and $e \in C$. Then by inductive hypothesis each element of G or $\{d\} \cup A$ has a rank-preserving grounding argument, so merge these arguments preserving rank-order, and add e to the end, so producing a rank-preserving grounding argument for e. Thus $E \in GExts(S)$. ∎

(22.22) COROLLARY. *$E \in GExts(S)$ iff $E = \Lambda(S, E)$.*

(22.23) THEOREM. *If every reason in $\mathcal{D}$ is finite, then $E \in FGExts(S)$ iff $E = \Lambda_\omega(S, E)$.*

PROOF. Suppose every reason in $\mathcal{D}$ is finite. By Corollary 22.20, we need only show that $\Lambda_\omega = E$ implies $E \in FGExts(S)$. Suppose $\Lambda_\omega = E$. We first show $\Lambda_\omega = \Lambda$. Suppose, by way of contradiction, that $\Lambda \neq \Lambda_\omega$. Then there must be a least ordinal $\alpha \geq \omega$ such that for some $e \in \mathcal{D}$, $e \in \Lambda_{\alpha+1} - \Lambda_\alpha$. Since α is minimal, $\Lambda_\omega = \Lambda_\alpha$, for otherwise $\Lambda_\omega = \Lambda_{\omega+1}$ and hence $\Lambda_\omega = \Lambda$. If $e \in \mathrm{Th}(\Lambda_\omega)$ then there is some $G \subseteq \Lambda_\omega$ such that $G \vdash \{e\}$. Since $\vdash$ is compact, there is a finite $G' \subseteq G$ such that $G' \vdash \{e\}$. But then the rank of G' is finite, say β, so $e \in \Lambda_{\beta+1}$, a contradiction. If $e \notin \mathrm{Th}(\Lambda_\omega)$, then by construction, there is some $f \in \Lambda_\omega$, $\mathcal{I}(f) = A \setminus\setminus B \mid\!\!\sim C$, $A \subseteq \Lambda_\omega$, $E \subseteq B^c$, and $e \in C$. Since A is finite, this means the rank of A is also finite. Thus there is some $\beta < \omega$ such that $A \subseteq \Lambda_\beta$ and $f \in \Lambda_\beta$, so $e \in \Lambda_{\beta+1} \subseteq \Lambda_\omega$, a contradiction. Thus $\Lambda = \Lambda_\omega$, and since $\Lambda_\omega = E$, by Theorem 22.21 E is a grounded extension of S. We see that E is finitely grounded by induction on rank. Clearly, if $e \in \Lambda_0$, then $e \in S$, hence $\{e\}$ is a rank-preserving grounding set. Now suppose the rank of e is $\alpha + 1 < \omega$. If $e \in \mathrm{Th}(\Lambda_\alpha)$, then there is a finite $G \subseteq \Lambda_\alpha$ such that $G \vdash \{e\}$. If $e \notin \mathrm{Th}(\Lambda_\alpha)$, then by construction there is some $f \in \Lambda_\alpha$ with $\mathcal{I}(f) = A \setminus\setminus B \mid\!\!\sim C$, $A \subseteq \Lambda_\alpha$, $E \subseteq B^c$, and $e \in C$. By inductive hypothesis, each of G or $f \cup A$ have finite rank-preserving grounding sets, so merge these preserving rank-order, add e to the end, and the result is a finite rank-preserving grounding order for e. ∎

(22.24) THEOREM (GROUNDED MINIMALITY). *$GExts \subseteq \mu QExts$.*

PROOF. $GExts \subseteq QExts$ by Theorem 22.9, so suppose $E \in GExts(S)$, $E' \in QExts(S)$, and $E' \subseteq E$. We first show $\Lambda(S, E) \subseteq \Lambda(S, E')$ by induction. Clearly $\Lambda_0(S, E) \subseteq \Lambda_0(S, E')$ since each equals S. Assume $\Lambda_\beta(S, E) \subseteq \Lambda_\beta(S, E')$ for each $\beta < \alpha$. If α is a limit ordinal, then by definition $\Lambda_\alpha(S, E') \subseteq \Lambda_\alpha(S, E)$. If α is a successor ordinal, say $\alpha = \beta + 1$, let $e \in \Lambda_\alpha(S, E)$. If $e \in S$ or $e \in \mathrm{Th}(\Lambda_\beta(S, E))$, then $e \in E'$, and otherwise there is a $d \in \Lambda_\beta(S, E)$ with $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\sim C$, $A \subseteq \Lambda_\beta(S, E)$, $E \subseteq B^c$, and $e \in C$. But since $E' \subseteq E \subseteq B^c$, this means $e \in \Lambda_\alpha(S, E')$. Hence $\Lambda(S, E) \subseteq \Lambda(S, E')$. But by Theorem 22.17 and Lemma 22.16, $E = \Lambda(S, E) \subseteq \Lambda(S, E') \subseteq E'$. ∎

(22.25) COROLLARY. *$GExts = \mu GExts$ and $FGExts = \mu FGExts$.*

(22.26) COROLLARY. *If $S \lhd \mathcal{D}$, then $AExts(S) = \{\mathcal{D}\}$.*

(22.27) COROLLARY. *If $S \lhd \mathcal{D}$, then $A \mid\!\models_S B$ for every $A, B \subseteq \mathcal{D}$.*

(22.28) COROLLARY. *If $A \mid\!\sim_S B$ for every $A, B \subseteq \mathcal{D}$, then $A \mid\!\models_S B$ for every $A, B \subseteq \mathcal{D}$.*

(22.29) COROLLARY. *Let $d, e \in \mathcal{D}$ be such that $\mathrm{Th}(\{d, e\}) = \mathcal{D}$. Then if both are cotenable in S, everything is arguable, that is $\emptyset \mid\!\sim_S B$ for every $B \subseteq \mathcal{D}$, and if both are inevitable in S, then everything is inevitable, in fact $A \mid\!\models_S B$ for every $A, B \subseteq \mathcal{D}$.*

PROOF. If $S \lhd E$ and $d, e \in E$, then $E = \mathcal{D}$, so by definition $\emptyset \mid\!\sim_S B$ for every $B \subseteq \mathcal{D}$. If S is incoherent, everything is inevitable by Theorem 14.1, and if S is coherent and both d and e are inevitable, they are cotenable, so $AExts(S) = \{\mathcal{D}\}$ and again Theorem 14.1 applies. ∎

This result holds special interest later when, in logical language domains, we consider state components $p, \neg p$ such that $\mathrm{Th}(\{p, \neg p\}) = \mathcal{D}$.

(22.30) DEFINITION. *The deductive reduction $(\mathcal{D}', \mathcal{I}', \vdash')$ of a deductively closed agent $(\mathcal{D}, \mathcal{I}, \vdash)$ is an agent such that (1) $\mathcal{D}' = \mathcal{D}$, (2) for each $d \in \mathcal{D}$, if $\mathcal{I}(d) = A \setminus\!\setminus \emptyset \mid\!\!\vdash C$, then $\mathcal{I}'(d) = \mathbf{P}\,\mathcal{D}$, otherwise $\mathcal{I}'(d) = \mathcal{I}(d)$, and (3) $\vdash'$ is the least deductive closure relation on $\mathcal{D}$ such that (a) $A \vdash' B$ whenever $A \vdash B$, and (b) for each $d \in \mathcal{D}$, if $\mathcal{I}(d) = A \setminus\!\setminus \emptyset \mid\!\!\vdash C$, then $\{d\} \cup A \vdash' C$.*

(22.31) THEOREM. *If $(\mathcal{D}, \mathcal{I}', \vdash')$ is the deductive reduction of $(\mathcal{D}, \mathcal{I}, \vdash)$, then $\mathscr{S}' = \mathscr{S}$, $\mathcal{J}' = \mathcal{J}$, and $AExts' = AExts$.*

PROOF. First note by definition that $\mathcal{I}' \supset \mathcal{I}$ and $\mathrm{Th}' \supset \mathrm{Th}$, so $\mathcal{Q}' \supset \mathcal{Q}$ and $\mathcal{R}' \subseteq \mathcal{R}$.

($\mathscr{S}' = \mathscr{S}$) Suppose $S \in \mathscr{S}'$. Then $S \in \mathcal{R}'$, so $S \in \mathcal{R}$. Now if $d \in S$ is non-monotonic, then $S \in \mathcal{I}(d)$ by definition, and if $\mathcal{I}(d) = A \setminus\!\setminus \emptyset \mid\!\!\vdash C$ and $A \subseteq S$, then $C \subseteq S$ by deductive closure, so $S \in \mathcal{I}(d)$. Thus $S \in \mathcal{Q}$, hence $S \in \mathscr{S}$. Now suppose $S \in \mathscr{S}$. Then $S \in \mathcal{Q}$, so $S \in \mathcal{Q}'$. Clearly $S \subseteq \mathrm{Th}'(S)$, yet if $A \vdash' B$ and $A \subseteq S$, then $A \vdash' B$ is generated by some deductions in $\vdash$ and some reasons. But since $S \in \mathcal{Q}$, these reasons must be satisfied, and produce the necessary conclusions, so $B \subseteq S$ as well. Thus $S \in \mathcal{R}'$, so $S \in \mathscr{S}'$.

($\mathcal{J}' = \mathcal{J}$) By the preceding, $Exts = Exts'$, so let $d \in \mathcal{D}$ and $S, E \subseteq \mathcal{D}$. If $d \in S$, then E satisfies both $\mathcal{J}(d, S)$ and $\mathcal{J}'(d, S)$, so suppose $d \notin S$. First, if $E \in \mathcal{J}(d, S)$, and if $E - \{d\} \vdash \{d\}$, then $E - \{d\} \vdash' \{d\}$, and if $E - \{d\} \nvdash' \{d\}$, then there is an $e \in E$, $\mathcal{I}(e) = A \setminus\!\setminus B \mid\!\!\vdash C$, $A \subseteq E \subseteq B^c$ and $d \in C$. If $B \neq \emptyset$, this e satisfies $\mathcal{J}'(d, S)$ too, and if $B = \emptyset$, then $E - \{d\} \vdash' \{d\}$. Hence $E \in \mathcal{J}'(d, S)$. Second, suppose $E \in \mathcal{J}'(d, S)$. If some valid reason supports d in E, it also does so for $\mathcal{J}(d, S)$. If no valid reason supports d in E, then $E - \{d\} \vdash' \{d\}$. In this case, if $E - \{d\} \nvdash \{d\}$, there is a monotonic reason which supports d for $\mathcal{J}(d, S)$. Hence $E \in \mathcal{J}(d, S)$, and $\mathcal{J}' = \mathcal{J}$.

($AExts' = AExts$) By the preceding, $QExts' = QExts$. First suppose $E \in FGExts(S)$. Then each $d \in E$ has a grounding argument G, which is also a grounding argument in the deductive reduction under reinterpretation of the steps. Thus $E \in FGExts'(S)$. Second, suppose $E \in FGExts'(S)$. Then each $d \in E$ has a grounding argument G. But the steps of G can each be reinterpreted as several steps in the unreduced agent, so G is also a grounding argument there. Thus $E \in FGExts(S)$, so $AExts' = AExts$. ∎

(22.32) COROLLARY (NON-MONOTONIC INCONSISTENCY IS SEMI-CLASSICAL (REITER)). *If $A \mid\!\sim_S B$ for every $A, B \subseteq \mathcal{D}$, then $\mathrm{Th}'(S) = \mathcal{D}$ in the deductive reduction of the agent.*

PROOF. Suppose $A \mid\!\sim_S B$ for every $A, B \subseteq \mathcal{D}$. By Theorem 14.2 this means $S \lhd \mathcal{D}$, hence $AExts(S) = \{\mathcal{D}\}$. By the previous theorem, $AExts'(S) = \{\mathcal{D}\}$, so $\mathcal{D} \in FGExts(S)$. Since no non-monotonic reasons can be valid in $\mathcal{D}$, by the definition of groundedness and the transitivity of deducibility, this means $S \vdash' \{d\}$ for each $d \in \mathcal{D}$, hence $\mathcal{D} = \mathrm{Th}'(S)$. ∎

§**23.** The *invertible reasons* theory is an augmentation of the deductively closed simple reasons theory in which all state components have negations, and in which contradicted conclusions can be passed backwards through reasons to contradict antecedents or raise qualifications. Intuitively, this theory recalls a logical structure for reasons in which $A \setminus\setminus B \mid\!\!\sim C$ really means $A \wedge \neg B \supset C$, so that when asserting $\neg C$, $\neg A \vee B$ can be inferred. Of course, $A \setminus\setminus B \mid\!\!\sim C$ cannot really be $A \wedge \neg B \supset C$, for then the normal default $\emptyset \setminus\setminus \neg C \mid\!\!\sim C$ becomes $\neg C \supset C$. This is equivalent to simply C, so that under this rewriting defaults are pure axioms rather than rules which may or may not introduce reasoned assumptions. We escape this absolute triviality by making the notion of admissible extension more complex, allowing individual reasons to be used in different ways in different circumstances. Instead of at most contributing its conclusions C, a reason $A \setminus\setminus B \mid\!\!\sim C$ in this theory can also be used to infer elements of $\neg A$ or B on occasion. Unfortunately, as we shall see soon, even this more involved reading of invertible reasons trivializes theories of reasoned assumptions, albeit in a different way.

(23.1) DEFINITION. *An agent's use of invertible reasons is characterized by $(\mathcal{D}, \mathcal{I}, \mathcal{R}, \mathcal{J}, \lhd)$ together with relations $\neg$ and $\vdash$ such that*

(i) $\mathcal{D}$ is a set,
(ii) $\neg : \mathcal{D} \to \mathcal{D}$ is 1-1,
(iii) $\vdash$ is a deducibility relation such that for each $d \in \mathcal{D}$, $[d] = [\neg\neg d]$,
(iv) For each $d \in \mathcal{D}$, there are sets $A, B, C \subseteq \mathcal{D}$ such that

$$\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\sim C = \{S \subseteq \mathcal{D} \mid A \subseteq S \subseteq B^c \supset C \subseteq S\}$$

(v) $\mathcal{R} = \{S \subseteq \mathcal{D} \mid S = \mathrm{Th}(S)\}$,
(vi) For each $d \in \mathcal{D}$ and $S \subseteq \mathcal{D}$,

$$\begin{aligned}\mathcal{J}(d,S) = \{E \quad | \quad & d \in S \ \vee\ E - \{d\} \vdash \{d\} \ \vee\ \exists e \in E\, \exists A, B, C \subseteq \mathcal{D} \\ & \mathcal{I}(e) = A \setminus\setminus B \mid\!\!\sim C \ \wedge \\ & [(d \in C \ \wedge\ A \subseteq E \subseteq B^c) \vee \\ & \ (d \in B \ \wedge\ A \subseteq E \ \wedge\ E \cap B = \{d\} \ \wedge\ E \cap \neg C \neq \emptyset) \vee \\ & \ (d = \neg d' \in \neg A \ \wedge\ A - E = \{d\} \ \wedge\ E \subseteq B^c \ \wedge\ E \cap \neg C \neq \emptyset)]\}\end{aligned}$$

(vii) $AExts = FGExts$, where $E \in FGExts(S)$ iff $S \lhd E$ and for each $e \in E$ there is a finite grounding set $G \subseteq E$ well-ordered by $<_G$ such that $e \in G$ and whenever $d \in G$, either (1) $d \in S$, or (2) there is some $f \in G$ and $A, B, C \subseteq \mathcal{D}$ such that $f <_G d$, $\mathcal{I}(f) = A \setminus\setminus B \mid\!\!\sim C$ and either (a) $d \in C$, $A \subseteq G$, $A <_G d$, and $E \subseteq B^c$, or (b) $d \in B$, $A \subseteq G$, $A <_G d$, $E \cap B = \{d\}$, and for some $c \in C$, both $\neg c \in G$ and $\neg c <_G d$, or (c) $d = \neg d' \in \neg A$, $A - E = \{d'\}$, $(A - E) <_G d$, $E \subseteq B^c$, and for some $c \in C$, both $\neg c \in G$ and $\neg c <_G d$, or (3) there is some $A <_G d$ such that $A \vdash \{d\}$.

The definition of groundedness means that every element in an admissible extension of S must either be an element of S itself, follow deductively from other elements of the admissible extension, or contribute to the satisfaction of some reason in the admissible extension, where the satisfaction of the reason and the circumstances of the other elements it mentions necessitate introducing the element in order to make the appropriate condition true or false. Unfortunately, invertibility guts the notion that reasons express preferences about what assumptions to make, leaving only statements of anti-agnosticism.

(23.2) THEOREM. *If $\mathcal{I}(a) = \{\emptyset \setminus\setminus \{\neg b\} \mid\!\!\sim \{b\}\}$ and $\mathcal{I}(b) = \emptyset \setminus\setminus \emptyset \mid\!\!\sim \emptyset$, then $AExts(\{a\}) = \{\mathrm{Th}(\{a, b\}),\quad \mathrm{Th}(\{a, \neg b\})\}$.*

PROOF. Each of these sets is clearly an extension of $\{a\}$, with $\mathrm{Th}(\{a, b\})$ an admissible extension as usual. $\mathrm{Th}(\{a, \neg b\})$ is also a finitely grounded extension, since $\neg b$ has the sequence $\langle a, \neg b\rangle$ which satisfies condition (2)(b) of the definition. ∎

.

(23.3) CONJECTURE (INVERTIBLE TRIVIALITY). *In the theory of invertible reasons, $AExts = \mu Exts$.*

§**24.** We now finally specialize the theory to one involving a logical language in the *linguistic reasons* theory. Where previously we have not cared how state components express reasons, the idea of the linguistic reasons theory is to express reasons in logical syntax. Since reasons refer to sets of state components, some of which may be other reasons, we must choose the language so that formulas may refer to sets of other formulas. The techniques for accomplishing this are well known, if tedious. It is sufficient for our purposes to restrict all sets of formulas to explicit presentations, to only finite sets, and no quantification over sets or the arguments of $\setminus\!\setminus\ \Vdash$ allowed. (One can, of course, allow such quantification and thus make the language at least weak second order.) In summary, we take the language $\mathcal{L}$ to contain the $\setminus\!\setminus\ \Vdash$ symbol and to be a metalanguage of itself using the quasi-quote notation $\ulcorner\ \urcorner$ for naming formulas and finite sets of formulas. A prime use of these expressions are in stating schematic or quantified reasons, for example

$$\forall x\,[\ulcorner\{p(x)\}\urcorner \setminus\!\setminus \ulcorner\{q(x)\}\urcorner \Vdash \ulcorner\{r(x)\}\urcorner\,].$$

Here an expression like $\ulcorner\{p(x)\}\urcorner$ represents an open term built up out of the name-constructing functions and the free variable x. For readability, we omit most explicit uses of quasi-quotes in the following since their use should be clear.

(24.1) DEFINITION. *An agent's use of linguistic reasons is characterized by the axioms for simple reasons in deductively closed states, where*

(i) $\mathcal{D}$ *is the set of sentences of a language* $\mathcal{L}$ *as above,*
(ii) $\vdash$ *is ordinary deducibility*
(iii) For each $d \in \mathcal{D}$*,*

$$\mathcal{I}(d) = \begin{cases} \{S \subseteq \mathcal{D} \mid A \subseteq S \subseteq B^c \supset C \subseteq S\} & \text{if } d \text{ is closed and } d = \ulcorner A\urcorner \setminus\!\setminus \ulcorner B\urcorner \Vdash \ulcorner C\urcorner\text{;} \\ \mathbf{P}\,\mathcal{D} & \text{otherwise.} \end{cases}$$

Since states are deductively closed sets of sentences and universally quantified formulas imply all their closed instances, quantified reasons are interpreted as shorthand for all their closed instances. Note that local groundedness becomes trivial in this theory. Because $[d] = [\neg\neg d]$, if S is a deductively closed set of sentences, $S - \{d\} \vdash \{d\}$ for each $d \in S$, so $QExts = Exts$.

EXAMPLES. In addition to our earlier conventions, we assume that P and Q are ordinary closed wffs of $\mathcal{L}$.

(24.2) $S = \{\emptyset \setminus\!\setminus \{\neg P\} \Vdash \{P\}\}, \qquad AExts(S) = \{\mathrm{Th}(S \cup \{P\})\}$

(24.3) $S = \{\emptyset \setminus\!\setminus \{P\} \Vdash \{P\}\}, \qquad AExts(S) = \emptyset$

(24.4) $S = \{\emptyset \setminus\!\setminus \{\neg P\} \Vdash \{P\},\quad \emptyset \setminus\!\setminus \{P\} \Vdash \{\neg P\}\},$
$AExts(S) = \{\mathrm{Th}(S \cup \{P\}),\quad \mathrm{Th}(S \cup \{\neg P\})\}$

(24.5) $S = \{\emptyset \setminus\!\setminus \{\neg P\} \Vdash \{P\},\quad \emptyset \setminus\!\setminus \{\neg Q\} \Vdash \{Q\},\quad \neg(P \wedge Q)\},$
$AExts(S) = \{\mathrm{Th}(S \cup \{P\}),\quad \mathrm{Th}(S \cup \{Q\})\}$

(24.6) $S = \{\emptyset \setminus\!\setminus A \Vdash \{P\},\quad \emptyset \setminus\!\setminus B \Vdash \{Q\},\quad \neg(P \wedge Q)\},$
$AExts(S) = \emptyset$

(24.7) $S = \{\forall n \in \mathbf{N}\,[\emptyset \setminus\!\setminus \{\neg p(n)\} \Vdash \{p(n)\}],\quad \forall n \in \mathbf{N}\; p(n+1) \supset \neg p(n)\},$
$AExts(S) = \{\mathrm{Th}(S \cup \{p(n) \mid n = 2m\}),\quad \mathrm{Th}(S \cup \{p(n) \mid n = 2m+1\})\}$

(24.8) $S = \{\forall n \in \mathbf{N}\,[\emptyset \setminus\!\setminus \{\neg p(n)\} \Vdash \{p(n)\}],$
$\forall n > 0\; p(n) \supset \exists m < n\; \neg p(m),\quad \forall n > 0\; \neg p(n) \supset \exists m < n\; p(m)\},$

This has an uncountable number of admissible extensions, one for each pattern of alternating p-ness and $\neg p$-ness. That is, for every sequence of positive integers $n_1, n_2, n_3, \ldots$ there is one admissible extension that contains

$$p(0), \ldots, p(n_1 - 1), \neg p(n_1), \ldots, \neg p(n_1 + n_2 - 1), p(n_1 + n_2), \ldots$$

and another admissible extension that contains the negations of all of these.

(24.9) $S = \{\exists x\, p(x),\quad \forall x[\emptyset \setminus\!\setminus \{p(x)\} \Vdash \{\neg p(x)\}\}$

This has a different admissible extension for each equivalence class $[p(t)]$ for closed terms t, namely

$$\mathrm{Th}(S \cup \{p(t)\} \cup \{\neg p(t') \mid t' \text{ a closed term and } [p(t')] \neq [p(t)]\}),$$

as well as an extension

$$\mathrm{Th}(S \cup \{\neg p(t) \mid t \text{ a closed term}\}),$$

in which the p-object has no name in the language.

(24.10) $S = \{\exists x\, p(x), \quad \forall x[\emptyset \setminus\setminus \{p(x)\} \mid\!\sim \{\neg p(x)\}, \quad \forall x(x = a \ \vee \ x = b)\}$
$AExts(S) = \{\mathrm{Th}(S \cup \{p(a), \neg p(b)\}), \quad \mathrm{Th}(S \cup \{\neg p(a), p(b)\})\}$

(24.11) $S = \{\forall n \in \mathbf{N}\,[\{lot(n+1)\} \setminus\setminus \{\neg lot(n)\} \mid\!\sim \{lot(n)\}],$
$\forall n \in \mathbf{N}\,[\{\neg lot(n)\} \setminus\setminus \{lot(n+1)\} \mid\!\sim \{\neg lot(n+1)\}],$
$\forall n \in \mathbf{N}\,[lot(n) \supset lot(n+1)],$
$\neg lot(0), \quad lot(3)\}$
$AExts(S) = \{\mathrm{Th}(S \cup \{lot(1), lot(2), \ldots\}),$
$\mathrm{Th}(S \cup \{\neg lot(1), lot(2), \ldots\}),$
$\mathrm{Th}(S \cup \{\neg lot(1), \neg lot(2), lot(3), \ldots\})\}$

The first axiom says that if $n+1$ is a lot, presumably so is n. The second says if n is few, presumably so is $n+1$. The third says that more than a lot is also a lot. Note that there is a distinct admissible extension for each choice of the boundary between a few and a lot.

(24.12) COROLLARY (NON-MONOTONIC INCONSISTENCY IS SEMI-CLASSICAL). *If S is inconsistent and contains no nontrivial monotonic reasons, S is classically inconsistent, that is,* $\mathrm{Th}(S) = \mathcal{L}$.

PROOF. This just restates Corollary 22.32 for the case of a deductively reduced linguistic reasons agent. ∎

(24.13) DEFINITION. *A normal reason is an element $d \in \mathcal{D}$ such that for some $A, C \subseteq \mathcal{D}$, $\mathcal{I}(d) = A \setminus\setminus \neg C \mid\!\sim C$.*

(24.14) THEOREM (ORTHOGONALITY OF EXTENSIONS (REITER)). *If $E, F \in AExts(S)$ contain only normal reasons and $E \neq F$, then $E \cup F$ is inconsistent.*

PROOF. Suppose $E, F \in AExts(S)$ contain only normal reasons, and that $E \neq F$. Now $E = \Lambda_\omega(S, E)$, $F = \Lambda_\omega(S, F)$, and $\Lambda_0(S, E) = \Lambda_0(S, F) = S$, so there must be a least $\alpha \geq 0$ such that $\Lambda_{\alpha+1}(S, E) \neq \Lambda_{\alpha+1}(S, F)$ while $\Lambda_\alpha(S, E) = \Lambda_\alpha(S, F)$. Without loss of generality, suppose $e \in \Lambda_{\alpha+1}(S, E) - \Lambda_{\alpha+1}(S, F)$. Then there is a $d \in \Lambda_\alpha$ such that $\mathcal{I}(d) = A \setminus\setminus \neg C \mid\!\sim C$, $A \subseteq \Lambda_\alpha$, $E \subseteq (\neg C)^c$, and $e \in C$. In fact, $C \subseteq E$. But since $e \notin \Lambda_{\alpha+1}(S, F)$, there is some $\neg f \in \neg C$ with $\neg f \in F$. Thus $f \in E$ and $\neg f \in F$, so $E \cup F$ is inconsistent. ∎

(24.15) THEOREM (NORMAL REASONS (REITER)). *If $\mathcal{L}$ is restricted so that all reasons are normal, then every $S \subseteq \mathcal{L}$ is coherent.*

PROOF. Suppose every reason is normal. If $\mathrm{Th}(S) = \mathcal{L}$, then $S \lhd \mathcal{L}$, so suppose $\mathrm{Th}(S) \neq \mathcal{L}$. We construct an extension E. Let $E_0 = S$, and for $i \geq 0$, let T_i be a maximal set of closed wffs such that (1) $E_i \cup T_i$ is consistent, and (2) if $u \in T_i$, then for some $e \in E_i$, $\mathcal{I}(e) = A \setminus\setminus \neg C \mid\!\sim C$, $u \in C$, $C \subseteq T_i$, and $A \subseteq E_i$. Define $E_{i+1} = \mathrm{Th}(E_i) \cup T_i$, and $E = \bigcup_{i=0}^{\infty} E_i$. We prove that for each $i \geq 0$, $E_i = \Lambda_i(S, E)$, hence $\Lambda_\omega(S, E) = E$, and $S \lhd E$.

Clearly, $E_0 = \Lambda_0 = S$, so assume $E_i = \Lambda_i$. We claim $E_{i+1} = \Lambda_{i+1}$.

($E_{i+1} \subseteq \Lambda_{i+1}$) Let $e \in E_{i+1}$. If $e \in \mathrm{Th}(E_i)$, then $e \in \Lambda_{i+1}$, and if $e \notin \mathrm{Th}(E_i)$ there is some $f \in E_i$, $\mathcal{I}(f) = A \setminus\setminus \neg C \mid\!\sim C$, $A \subseteq E_i$, $f \in C$, and $C \subseteq T_i$. But since $E_{i+1} \subseteq E$ is consistent, $E \subseteq (\neg C)^c$, hence $C \subseteq \Lambda_{i+1}$, so $e \in E_{i+1}$.

($\Lambda_{i+1} \subseteq E_{i+1}$) Let $e \in \Lambda_{i+1}$. If $e \in \mathrm{Th}(\Lambda_i)$, then $e \in E_{i+1}$, and if $e \notin \mathrm{Th}(\Lambda_i)$ then there is an $f \in \Lambda_i$, $\mathcal{I}(f) = A \setminus\setminus \neg C \mid\!\sim C$, $A \subseteq \Lambda_i$, $E \subseteq (\neg C)^c$, $C \subseteq \Lambda_{i+1}$, and $e \in C$. Now if $e \notin E_{i+1}$, by the maximality of T_i we know $E_i \cup T_i \cup \{e\}$ is inconsistent, so $\mathrm{Th}(E_i) \cup T_i \cup \{e\}$ is inconsistent, so $E_{i+1} \cup \{e\}$ is inconsistent. But since $E_{i+1} \subseteq E$, this means $E \cup \{e\}$ is inconsistent. Since clearly $E = \mathrm{Th}(E)$, we must have $\neg e \in E$, contradicting $E \subseteq (\neg C)^c$. Hence $e \in E_{i+1}$, so $\Lambda_{i+1} \subseteq E_{i+1}$. Thus $\Lambda_{i+1} = E_{i+1}$ and $S \lhd E$. ∎

Results are also known concerning the decidability of coherence, arguability, and inevitability in the linguistic reasons theory, but they are less satisfying than those of the simple reasons theory. Where previously a finite universe ensured computability, the connection inherent in this theory between arguability and logical consistency puts most problems beyond the bounds of recursiveness.

(24.16) THEOREM (REITER). $\bigcup AExts(S)$ *is not recursively enumerable in* S.

(24.17) COROLLARY. *Coherence is not decidable.*

(24.18) THEOREM (CHURCH). *Consistency is not decidable.*

For some special cases, such as finite sentential and monadic sets, arguability and inevitability appear decidable. But rather than continue this topic here, we refer to the discussions in [REITER 1980], [MCDERMOTT AND DOYLE 1980], and [DAVIS 1980].

§**25.** In the previous theories of reasoned assumptions, there was no commitment to what state components signified, other than reasons. Elements of the domain were not in themselves any familiar psychological organization, but merely the components from which to build mental states. As building blocks, the statements of logical languages can be used to encode versions of beliefs, desires, and other common psychological notions, but they are not beliefs or desires by themselves without further constitutive assumptions. It is tempting, of course, to phrase attitudinal theories directly in terms of logical theories, for attitudes are often taken to be attitudes towards propositions, and languages are the usual way of expressing propositions. However, such direct phrasing can hide fundamental ideas about attitudes among masses of particular linguistic details. To seek clarity, we temporarily retreat from logical theories, and work back up to logical forms after starting once again with the simple reasons theory.

§**26.** Pick an attitudinal ontology for a psychology, and we can cast its elements directly in the simple reasons theory. In the following, we assume for simplicity that the agent's attitudes divide into three classes: beliefs (**Bel**), desires (**Des**), and intentions (**Int**). We also assume each particular attitude among these classes is a possible state component, that is, $\mathbf{Bel}, \mathbf{Des}, \mathbf{Int} \subseteq \mathcal{D}$. This leaves open the possibility that $\mathcal{D}$ may contain non-attitudinal components. For example, we can take the set of reasons (**Rsn**) to be outside the sets of attitudes. In such a formulation, we make explicit in reasons the connections between the agent's attitudes, but we need not assign special import to the attitudes themselves. That is, we might assume that $\mathcal{I}(a) = \mathbf{P}\,\mathcal{D}$ for each $a \in \mathbf{Bel} \cup \mathbf{Des} \cup \mathbf{Int}$. Of course, in an ecological theory of the agent we would wish to interpret these attitudes substantively, so that beliefs, say, indicate sets of possible worlds in which they are true. But here our concern is with narrow theories of the agent, and we can ignore all substantive interpretations except those of reasons.

§**27.** Rather than resting content with the simple attitudinal theory, we can define further theories of reasoned assumptions by stipulating that states are composed exclusively of attitudes. In terms of the sets we introduced earlier, we assume $\mathcal{D} = \mathbf{Bel} \oplus \mathbf{Des} \oplus \mathbf{Int}$. But since reasons are state components, that is, $\mathbf{Rsn} \subseteq \mathcal{D}$, we must also say how reasons appear among the attitudes. Our initial motivations and formulations of reasons suggest two possibilities, namely $\mathbf{Rsn} \subseteq \mathbf{Int}$ and $\mathbf{Rsn} \subseteq \mathbf{Des}$. (The possibility $\mathbf{Rsn} \subseteq \mathbf{Bel}$ is not considered because our theories take reasons to be prescriptive, rather than merely descriptive.) Reasons act as specifications to be satisfied by states containing them, and such definite specifications might be thought of as intentions of the agent about its own construction. Taking $\mathbf{Rsn} \subseteq \mathbf{Int}$ produces what we can call the *ratiocinative intention* theory of reasoned assumptions. In this theory, we assume that if the agent has a state at all, it has satisfied all the intentions concerning that state expressed in that state, although we do not require that in achieving that state it satisfied earlier intentions about what to do. We produce yet another theory, that of *ratiocinative desires*, by assuming $\mathbf{Rsn} \subseteq \mathbf{Des}$. In this theory, we interpret reasons not as definite specifications for mental states, but as preferences for states satisfying certain conditions over states not satisfying those conditions. Here we take preferences to be desires to attain one alternative in every situation presenting a specified set of alternatives. How do we take such desires to be satisfied? Consider states in which as many ratiocinative desires are satisfied as possible, that is, states in which changes that would satisfy some unsatisfied desire would result in failures to satisfy some currently satisfied desire. (Notions much like this are familiar in economics as Pareto optimality, but pursuit of this connection here would digress too far.) It turns out that these states of maximal utility are exactly the admissible states as we have defined them. We see this formally as follows.

(27.1) DEFINITION. *A set $S \subseteq \mathcal{D}$ is satisfaction optimal iff for each $d \in S$, if $S \notin \mathcal{I}(d)$ then for each $S' \in \mathcal{I}(d)$ there is some $d' \in S$ with $S \in \mathcal{I}(d')$ but $S' \notin \mathcal{I}(d')$.*

(27.2) THEOREM. *If S is admissible, S is satisfaction optimal.*

PROOF. Trivially, since if S is admissible, there are no elements $d \in S$ such that $S \notin \mathcal{I}(d)$. ∎

(27.3) THEOREM. *There is an inadmissible satisfaction optimal set.*

PROOF. Let $\mathcal{D} = \{d, e\}$, $S = \{d\}$, $\mathcal{I}(d) = \{S\}$, and $\mathcal{I}(e) = \{\mathcal{D}\}$. Then $\mathcal{D}$ is inadmissible because $d \in \mathcal{D} \notin \mathcal{I}(d)$. But $\mathcal{D}$ is satisfaction optimal since d is the only unsatisfied element in $\mathcal{D}$, S is the only set in $\mathcal{I}(d)$, and $e \in \mathcal{D}$, $\mathcal{D} \in \mathcal{I}(e)$, and $S \notin \mathcal{I}(e)$. ∎

(27.4) THEOREM. *In the simple reasons theory, if S is satisfaction optimal, S is admissible.*

PROOF. We prove the contrapositive. Suppose $S \subseteq \mathcal{D}$ is inadmissible, that is, for some $d \in \mathcal{D}$, $d \in S \notin \mathcal{I}(d)$. Now by Theorem 20.1, $\mathcal{D} \in \mathcal{I}(d)$, but for all $e \in S$, $\mathcal{D} \in \mathcal{I}(e)$ as well, so S is not satisfaction optimal. ∎

(27.5) COROLLARY. *In the simple reasons theory, sets are admissible iff they are satisfaction optimal.*

(27.6) QUESTION. *Can interesting parts of the general theory be developed in terms of satisfaction optimality rather than admissibility as we have defined it? Or is the simple reasons theory the weakest theory of any interest?*

Another notion within the simple reasons theory similar to satisfaction optimality is validity optimality. This stems from a focus on ratiocinative desires which manage to make their objective inferences instead of on the mere satisfaction of specifications by either success or disqualification. We first recall the definition of validity.

(27.7) DEFINITION. *An element $d \in S$ is valid in S iff $\mathcal{I}(d) = A \setminus\setminus B \mid\!\!\!\vdash C$ and $A \subseteq S \subseteq B^c$. Otherwise, d is invalid in S. We write $V(d)$ to mean the set of all $S \subseteq \mathcal{D}$ such that d is valid in S. We write $\mathcal{V}(S)$ to mean the set of all $d \in S$ such that d is valid in S, and $\bar{\mathcal{V}}(S)$ to mean $S - \mathcal{V}(S)$, the set of all d invalid in S.*

Note that even invalid elements can be satisfied.

(27.8) DEFINITION. *A set S is validity optimal in $\hat{S} \subseteq \mathbf{P}\,\mathcal{D}$ iff for each $d \in \mathcal{D}$, if $S' \in \hat{S}$ validates d but S does not, then there is some e valid in S but not valid in S'. Alternatively, S is validity optimal in $\hat{S}$ iff for each $d \in S$, if $S \notin V(d)$ then for each $S' \in V(d) \cap \hat{S}$ there is some $d' \in S$ with $S \in V(d)$ but $S' \notin V(d')$.*

(27.9) COROLLARY. *If $\hat{S}_1 \supset \hat{S}_2$ and S is validity optimal in $\hat{S}_1$, then S is validity optimal in $\hat{S}_2$.*

(27.10) THEOREM. *There are admissible sets not validity optimal in $\mathscr{S}$.*

PROOF. Let $\mathcal{D} = \{d\}$, and $\mathcal{I}(d) = \mathbf{P}\,\mathcal{D}$. Then both $\emptyset$ and $\mathcal{D}$ are admissible, $\mathcal{D}$ validates d, but $\emptyset$ validates no element not in $\mathcal{D}$, so $\emptyset$ is not validity optimal in $\mathscr{S}$. ∎

(27.11) THEOREM. *There are inadmissible sets validity optimal in $\mathbf{P}\,\mathcal{D}$.*

PROOF. Let $\mathcal{D} = \{d, e\}$, $\mathcal{I}(d) = \emptyset \setminus\setminus \{d\} \mid\!\!\!\vdash \{e\}$, and $\mathcal{I}(e) = \emptyset \setminus\setminus \{e\} \mid\!\!\!\vdash \{d\}$. Then $\mathscr{S} = \{\emptyset, \mathcal{D}\} \neq \mathbf{P}\,\mathcal{D}$, but every set in $\mathbf{P}\,\mathcal{D}$ is validity optimal, since no set validates any element. ∎

Our intuition is that validity optimality captures the fundamentals of the orthogonality of admissible extensions of a set, of what appears in the linguistic reasons theory as inconsistency of alternative admissible extensions of sets of normal defaults. That is, this notion shows we can sensibly speak of psychological incompatibility without requiring notions of logical inconsistency.

(27.12) THEOREM. *In the simple reasons theory, if $S \lhd E$, then E is validity optimal in $AExts(S)$.*

PROOF. Suppose, by way of contradiction, that $E \in AExts(S)$ and that there is $E' \in AExts(S)$ such that d is valid in E' but not valid in E, yet no e valid in E is not valid in E'. That is, there is a $d \in E'$, d valid in E and either $d \notin E$ or d invalid in E', and $\mathcal{V}(E) \subseteq \mathcal{V}(E')$. The element d shows $E \neq E'$, so by the minimality of E and E' among $QExts(S)$, $E - E' \neq \emptyset$. Let $e \in E - E'$. Since $S \subseteq E'$, $e \notin S$. But since $E \in QExts(S)$, there must be some $f \in E$, f valid in E, and e a consequence of f. But then f is valid in E' as well, so $e \in E'$, a contradiction. Thus E must be validity optimal in $AExts(S)$. ∎

(27.13) QUESTION. *If E is validity optimal in $AExts(S)$, is $E \in AExts(S)$?*

(27.14) DEFINITION. *A set S is strongly validity optimal in $\hat{S} \subseteq \mathbf{P}\,\mathcal{D}$ iff for each $d \in \mathcal{D}$, if $S' \in \hat{S}$ validates d while S invalidates d, then there is some e valid in S but invalid in S'; in other words, if $\mathcal{V}(S') \cap \bar{\mathcal{V}}(S) \neq \emptyset$, then $\bar{\mathcal{V}}(S') \cap \mathcal{V}(S) \neq \emptyset$.*

(27.15) COROLLARY. *If S is strongly validity optimal in $\hat{S}$, then S is validity optimal in $\hat{S}$.*

(27.16) THEOREM. *If $S \lhd E$ in a finite simple reasons agent, then E is strongly validity optimal in $AExts(S)$.*

PROOF. Suppose $S \lhd E$ in a finite simple reasons agent and E is not strongly validity optimal. Then there is $E' \in AExts(S)$ and $d \in E$ such that d is valid in E' but invalid in E, and if e is valid in E, then either $e \notin E'$ or e is valid in E' as well. The differing properties of d show $E \neq E'$, so by the minimality of E and E' in $QExts(S)$, $E - E' \neq \emptyset$. Now $E = \Lambda_\omega(S, E)$, $E' = \Lambda_\omega(S, E')$, and $S = \Lambda_0(S, E) = \Lambda_0(S, E')$, so there is a least $\alpha \geq 0$ such that $\Lambda_{\alpha+1}(S, E) \neq \Lambda_{\alpha+1}(S, E')$ but $\Lambda_\alpha(S, E) = \Lambda_\alpha(S, E')$. Without loss of generality, suppose $e \in \Lambda_{\alpha+1}(S, E) - \Lambda_{\alpha+1}(S, E')$. Then there is some $f \in \Lambda_\alpha$ with $\mathcal{I}(f) = A \setminus\setminus B \mid\!\!\sim C$, $A \subseteq \Lambda_\alpha$, $E \subseteq B^c$, and $e \in C$. Since $e \notin \Lambda_{\alpha+1}(S, E')$, there must be some $g \in E' \cap B$, so f is invalid in E'. This contradicts the previous conclusion that since f is valid in E, either $f \notin E'$ or f is valid in E'. Hence E must be strongly validity optimal in $AExts(S)$. ∎

§**28.** As promised, we now consider propositional attitude theories of reasoned assumptions. To do this, we simply take the linguistic reasons theory and assume among the unary predicates of its language are symbols **Bel**, **Des**, and **Int**. Thus the classes **Bel**, **Des**, and **Int** are the ground wffs of the form, respectively, **Bel**$(\cdot)$, **Des**$(\cdot)$, and **Int**$(\cdot)$. The content of these attitudes is expressed by naming formulas of the language, for instance **Bel**$(\ulcorner 2 + 2 = 4 \urcorner)$. The theories of ratiocinative intentions and desires can then be had by only allowing reasons to occur as intentions or desires, respectively, as **Int**$(\ulcorner A \setminus\setminus B \mid\!\!\sim C \urcorner)$ or **Des**$(\ulcorner A \setminus\setminus B \mid\!\!\sim C \urcorner)$. That is, we now count expressions of the form $A \setminus\setminus B \mid\!\!\sim C$ as terms rather than as formulas, so that no sentence has the form $A \setminus\setminus B \mid\!\!\sim C$. To make up for this we interpret these particular sorts of intentions and desires specially, respectively by

$$\mathcal{I}(\textbf{Int}(\ulcorner A \setminus\setminus B \mid\!\!\sim C \urcorner)) = \{S \subseteq \mathcal{D} \mid A \subseteq S \subseteq B^c \supset C \subseteq S\}$$

$$\mathcal{I}(\textbf{Des}(\ulcorner A \setminus\setminus B \mid\!\!\sim C \urcorner)) = \{S \subseteq \mathcal{D} \mid A \subseteq S \subseteq B^c \supset C \subseteq S\}.$$

With these encodings, we can consider three new theories. The first is the *deductively closed beliefs* theory, in which we require the set of believed formulas to be closed under the deducibility relation $\vdash$. That is, we take

$$\mathcal{R} = \{S \subseteq \mathcal{D} \mid S = \mathrm{Th}(S) \ \wedge\ \exists S' \subseteq \mathcal{D} \quad [S' = \mathrm{Th}(S') \ \wedge\ \forall d \in \mathcal{D} \quad [d \in S' \equiv \textbf{Bel}(\ulcorner d \urcorner) \in S]]\}.$$

The second theory involving propositional attitudes is the *self-omniscient* theory, in which we assume the agent has complete and correct beliefs about its own attitudes. That is, we take

$$\mathcal{R} = \{S \subseteq \mathcal{D} \mid S = \mathrm{Th}(S) \ \wedge\ \forall d \in \mathcal{D} \quad [d \in S \equiv \textbf{Bel}(\ulcorner d \urcorner) \in S \ \wedge\ d \notin S \equiv \neg\textbf{Bel}(\ulcorner d \urcorner) \in S]\}.$$

Naturally, if $S \in \mathcal{R}$ is nonempty, it is infinite, containing reflection upon reflection of its own contents.

The third theory involving propositional attitudes combines these two into the *deductively self-omniscient* theory. Unfortunately, if $\vdash$ is the ordinary deducibility relation of logic, the theory is inconsistent, that is, there are no agents whose states satisfy all these requirements. Proving this here would digress too far from our main concerns. MONTAGUE proves it in some generality, and THOMASON discusses its significance for attempts to define notions of "semantic competence" of agents, notions corresponding to those of "grammatical competence" common in linguistics.[15]

[15] [MONTAGUE 1963], [THOMASON 1979]

§**29.** Our initial discussion of problems of acting with incomplete information focussed on adopting and abandoning assumptions, but the preceding sections made little mention of such activities. Rather, the task was to set out the admissible states of the agent, so that actions in which the agent changes its state may be better understood. With a wide assortment of structures for states now available, we return to the question of state changes and actions.

The notion of reasoned assumptions was developed to treat reasoned adoption and reasoned abandonment of assumptions. In reasoned adoption of assumptions, the agent acts to adopt a new assumption by adding a reason to its state whose conclusion is the new assumption and whose antecedents and qualifiers may indicate the considerations involved in the decision to adopt the assumption. But just adding a reason to an admissible state will not in general yield another admissible state, so we must find some other way of effecting the change. To do this, we interpret the new reason as the agent's partial specification of its next state. We may not wish to accept just any new state satisfying the reason, for such might abandon all of the previous state. If not, we can employ the idea of conservation and require that the new state should be an admissible state satisfying the new specification, one as "close" as possible to the previous state. Similarly, in reasoned abandonment of assumptions, the agent acts to rid itself of some unwanted assumption. This can be done either by removing from the state the reason supporting the assumption, or by adding to the state a new reason which defeats the assumption. But as before, the removal of a reason from an admissible state may not yield another admissible state, so we again may employ conservation in moving to a new admissible state which satisfies the new reason.

However useful might be reasoned adoption and abandonment of assumptions, one cannot always escape the need for forced or unreasoned adoption and abandonment of state components. The notion of forced adoption of state components might be used as a very crude way of viewing some effects of non-mental parts of a human on its state of mind, for example changes made by the perceptual and motor systems. Other effects, such as physiological changes in the body which affect the nervous system, seem ill-suited to this view, and require some other means of formulation, conceivably that of state-dependent interpretations, although I suspect other ideas are necessary as well. We do not explicitly treat forced adoption or abandonment of assumptions in this paper. On the other hand, forced changes of assumptions have a long history in artificial intelligence as the technique of backtracking, and we give a straightforward treatment of them later.

§**30.** Let $(\mathcal{D}, \mathcal{I}, \mathscr{S}, \mathscr{J}, \lhd)$ describe an agent and let $\mathcal{E}$ be a set of "environments." We suppose the basis for the agent's state changes to be described by a "kernel" transition function $\partial : \mathscr{S} \times \mathcal{E} \to \mathbf{P}\,\mathcal{D}$ which for each admissible state and environment yields the (partial) specifications for the next state. The role of this kernel as specifications for the next state is ensured by requiring the next state to be an extension of the kernel, since extensions of a set satisfy all the specifications represented by that set. In fact, we require $\Delta(S, e)$, the *admissible transitions from* S *in* e, to lie among the admissible extensions of the kernel, that is, $\Delta(S, e) \subseteq AExts(\partial(S, e))$. As with the notions of admissible states and admissible extensions, we capture different agents by stipulating Δ as different restrictions of $AExts \circ \partial$. However, while the earlier sorts of restrictions separated into local and global restrictions, here we only treat global restrictions on transitions, as formalization of interesting local restrictions involves other notions beyond the scope of this paper.

§**31.** Of possible sorts of global restrictions on admissible transitions, three occupy our attention in the following: those of strictness, of conservatism, and of their combination, strict conservatism. In spite of such labels, these notions have nothing to do with New Englanders.

Strictness we have seen previously in the form of strict arguability. In a strict agent, $\Delta = \mu AExts \circ \partial$. As usual, this is just $AExts \circ \partial$ if the previous theories are used to supply $AExts$, since in those $\mu AExts = AExts$.

Strictness requires that successor states stay as close as possible to the specifications of the kernel, giving up any part of the previous state not derivable from the kernel alone. This sort of agent captures the operation of RMS quite well, as explained in §46, in which the agent maintains a set of fundamental or kernel reasons, and regularly reconstructs the complete state when this kernel is modified by additions or emendations.

Where strictness requires successor states to be as close as possible to the kernel specifications, *conservatism* requires successor states to approximate their predecessors as closely as possible. Of course, different notions of approximation are of interest in different circumstances, so we require only certain abstract properties of approximation comparisons to define conservatism. We express approximation comparisons in terms of a relation $\preceq$ on transitions, a quasi-order whose minimal elements include the null transitions. That is, we read $(S_1, S_2) \preceq (S_3, S_4)$ as saying that the transition from S_1 to S_2 is not larger than the transition from S_3 to S_4, and insist $\preceq$ be reflexive, transitive, and order no transition properly smaller than a null transition. Formally, $\preceq$ is a relation on $(\mathscr{S} \times \mathscr{S}) \times (\mathscr{S} \times \mathscr{S})$ such that for all admissible states,

(i) $(S, S') \preceq (S, S')$,
(ii) If $(S_1, S_1') \preceq (S_2, S_2')$ and $(S_2, S_2') \preceq (S_3, S_3')$, then $(S_1, S_1') \preceq (S_3, S_3')$,
(iii) $(S, S) \preceq (S, S')$,
(iv) $(S', S') \preceq (S, S')$.

We define the admissible transitions of conservative agents in terms of $\preceq$-minimal transitions. In analogy with μ, we use the operator ν ("nearest") to indicate transition minimization. If $S \in \mathscr{S}$ and $X \subseteq \mathscr{S}$, we define

$$\nu(S, X) = \{x \in X \mid \forall y \in X \quad (S, y) \preceq (S, x) \supset (S, x) \preceq (S, y)\},$$

and abuse the notation by writing νf as shorthand for $\lambda S.\nu(S, f(S))$, so that $\Delta = \nu AExts \circ \partial$ means that $\Delta(S, e) = \nu(S, AExts(\partial(S, e)))$ for each $S \in \mathscr{S}$ and $e \in \mathcal{E}$. As with μ, $\nu(S, X) = \emptyset$ if $X = \emptyset$, so $\nu(S, AExts(\partial(S, e))) = \emptyset$ if $AExts(\partial(S, e)) = \emptyset$.

Several comments on this definition are in order. First, we only require $\preceq$ to be a quasi-order rather than a partial order. That is, we do not stipulate antisymmetry, so that $x \preceq y$ and $y \preceq x$ need not imply $x = y$. This is made clear in the examples to follow. Second, as in partial orders, it can happen that both $x \not\preceq y$ and $y \not\preceq x$. In our usage, if $AExts(\partial(S, e))$ contains two incomparable states but no common smaller state, then $\Delta(S, e)$ will contain both states, and the system $(\mathscr{S}, \mathcal{E}, \Delta)$ will be nondeterministic. This too is illustrated below. Third, minimal mutilation has been formalized in the philosophical study of counterfactual conditionals using the notion of comparative similarity relations on states.[16] Here we note only that every order $\preceq$ on transitions induces a (not necessarily interesting) comparative similarity relation on $\mathscr{S}$, and *vice versa*. Read $X \overset{\prec}{}_{Z} Y$ as "X is as close to Z as is Y." The connection between these relations then obtains by defining

$$(S_1, S_2) \preceq (S_3, S_4) \text{ iff } S_1 = S_3 \text{ and } S_2 \overset{\prec}{}_{S_1} S_4 \text{ or } S_2 = S_4 \text{ and } S_1 \overset{\prec}{}_{S_2} S_3.$$

Conservative agents need not pass along sets of fundamental reasons in ∂, but can instead simply specify changes and let the $\preceq$-minimization conserve as much of the previous state as possible. In agents whose admissible extensions are not strict, for example locally grounded agents ($AExts = QExts$), this can lead to retention of mutually supportive but ungrounded complexes of elements in the successor state. Indeed, many discussions of belief revision in the philosophical literature propose versions of conservative agents, as do proponents of the deductivist approach to artificial intelligence. In this approach, $\mathcal{D}$ is the set of sentences of a first-order logical language, $\mathscr{S}$ is the set of all deductively closed and consistent sets of sentences, and admissible extensions are just minimal extensions (and so just deductive closures if consistent), that is, $AExts = \mu Exts$. Here ∂ produces "inputs" to or observations by the agent, new sentences possibly contradicting current beliefs, and Δ moves the agent to some extension of $\partial(S, e)$ as close as possible (by some measure) to S, for example, to $\mathrm{Th}(A \cup \partial(S, e))$ for some maximal subset A of S consistent with $\partial(S, e)$.

Finally, one can combine the notions of strictness and conservatism into *strict conservatism*. Here we define the admissible transitions by $\Delta = \nu\mu AExts \circ \partial$. This sort of agent moves to admissible extensions of the kernel which first are closest to the kernel, and second are closest to the preceding state.

[16] See [LEWIS 1973], [TURNER 1981].

Each of these sorts of agents calls for study, as they may be appropriate for different applications. Systems doing more or less thorough analysis of domains small in comparison to the whole world (such as electronic circuit analysis and other expert tasks) might always require strictness, or better yet, strict conservatism. Systems whose focus ranges over the whole world (general agents and natural language users) probably have too many uncertain interrelated assumptions to think of any subset as "the axioms," and so must settle for simple conservatism in spite of the logical *faux pas* this sometimes entails in those special cases in which the axioms are distinguished. These restrictions might be combined in other ways as well. Artificial intelligence systems often divide into many small subsystems. One might organize the agent so that each small subsystem is strictly conservative, while the large collection as a whole is merely conservative.

§**32.** Let us look at an example of a strictly conservative agent. For simplicity, we assume $\mathcal{E} = \{e\}$ and omit all mention of this constant environment. We begin with an agent described by the simple reasons theory, and want some notion of relative size of transitions. Since states are sets of state components, perhaps the simplest indication of the amount of change involved in going from state S_1 to state S_2 is the symmetric difference $S_1 \triangle S_2 = (S_1 - S_2) \cup (S_2 - S_1) = (S_1 \cup S_2) - (S_1 \cap S_2)$. The set $S_1 - S_2$ reveals how much of the initial state is lost in the new state, and the set $S_2 - S_1$, shows how much of the new state is gained relative to the old. Together these sets indicate the degree of conservation (or lack of it) involved in the change. Thus one way of comparing two transitions is to define $(S_1, S_2) \preceq (S_3, S_4)$ iff $S_1 \triangle S_2 \subseteq S_3 \triangle S_4$. Since the symmetric difference of any set with itself is empty, and since $\subseteq$ is a quasi-order on sets with $\emptyset$ as its minimum element, this definition yields the required properties of $\preceq$.

Now suppose the agent starts in the empty state $S_0 = \emptyset$ and $\partial(S_0) = \{\emptyset \setminus\setminus A \mid\!\sim B, \quad \emptyset \setminus\setminus C \mid\!\sim D\}$. Properly speaking, of course, this is $\mathcal{I}^*(\partial(S_0))$, but we drop the interpretation notation without confusion here. $\partial(S_0)$ has a single admissible extension, so $S_1 = \partial(S_0) \cup B \cup D$, the conservation condition being satisfied trivially. Suppose next that $\partial(S_1) = \partial(S_0) \cup \{\emptyset \setminus\setminus B \mid\!\sim A, \quad \emptyset \setminus\setminus D \mid\!\sim C\}$. By itself, this kernel is very ambiguous, with the four admissible extensions $\partial(S_1) + A \cup C,\ A \cup D,\ B \cup C,\ B \cup D$. But of these, one is closest to S_1, namely $S_2 = \partial(S_1) \cup B \cup D$. Next, suppose $\partial(S_2) = \partial(S_1) \cup \{\emptyset \setminus\setminus E \mid\!\sim C\}$. Now the kernel has only two admissible extensions, namely $\partial(S_2) + A \cup C,\ B \cup C$. But of these, again one is closer to S_2 than the other, so $S_3 = \partial(S_2) \cup B \cup C$. Finally, suppose $\partial(S_3) = \partial(S_2) \cup \{\emptyset \setminus\setminus F \mid\!\sim G, \quad \emptyset \setminus\setminus G \mid\!\sim F\}$. There are now again four admissible extensions of the kernel, $\partial(S_3) + A \cup C \cup F,\ A \cup C \cup G,\ B \cup C \cup F,\ B \cup C \cup G$, and of these, two are incomparable nearest neighbors of S_3, namely $\partial(S_3) \cup B \cup C \cup F$ and $\partial(S_3) \cup B \cup C \cup G$. Thus $\Delta(S_3)$ contains both of these states.

If this agent were merely strict, rather than strictly conservative, the admissible transitions would be different. For instance, S_1 would remain the same, since it is the only admissible extension of $\partial(S_0)$, but S_2 could be any of the four admissible extensions of $\partial(S_1)$, since the transition is not required to be $\preceq$-minimal.

To exhibit the difference between conservative and strictly conservative agents, suppose that $S = \{a, b, c, d, e\}$, where $\mathcal{I}(a) = \{c\} \setminus\setminus \emptyset \mid\!\sim \{d\}$, $\mathcal{I}(b) = \{d\} \setminus\setminus \emptyset \mid\!\sim \{c\}$, $\mathcal{I}(c) = \mathcal{I}(d) = \emptyset \setminus\setminus \emptyset \mid\!\sim \emptyset$, and $\mathcal{I}(e) = \emptyset \setminus\setminus \emptyset \mid\!\sim \{c\}$. According to the simple reasons theory, S is admissible, and is an admissible extension of its subset $\{a, b, e\}$. Now suppose $\partial(S) = \{a, b\}$. This has two component-admissible extensions in S, namely $\{a, b\}$ and $\{a, b, c, d\}$. Since the former is grounded, it is the only admissible extension of $\partial(S)$ in the strict simple reasons theory. However, the latter extension is closer to S, and so would be chosen by the symmetric difference relation if admissible extensions were not required to be grounded.

Besides this notion of closeness based on symmetric difference, there are many others that are more appropriate in other circumstances. For example, one might define one transition smaller than another if the components conserved in the first include those conserved in the second, that is $(S_1, S_2) \preceq (S_3, S_4)$ iff $S_1 \cap S_2 \supset S_3 \cap S_4$. Or instead of considering sets, one might use their cardinalities, as in $(S_1, S_2) \preceq (S_3, S_4)$ iff $|S_1 \triangle S_2| \leq |S_3 \triangle S_4|$. There are many other possibilities for closeness relations based on tolerances, topologies, metrics, and measures, but we cannot treat these here.[17]

[17] See [QUINE 1953], [RESCHER 1964], and [GÄRDENFORS 1980].

§**33.** Let us summarize the discussion.

(33.1) DEFINITION. *Plain, strict, conservative, and strictly conservative agents are characterized by* $(\mathcal{D}, \mathcal{I}, \mathscr{S}, \mathcal{J}, \lhd)$ *as before, together with a set of environments* $\mathcal{E}$*, a kernel transition function* $\partial : \mathscr{S} \times \mathcal{E} \to \mathbf{P}\,\mathcal{D}$*, and an admissible transition table* $\Delta : \mathscr{S} \times \mathcal{E} \to \mathbf{P}\,\mathscr{S}$ *such that for each* $S \in \mathscr{S}$ *and* $e \in \mathcal{E}$*,* $\Delta(S, e) \subseteq AExts(\partial(S, e))$*. In a plain agent,* $\Delta = AExts \circ \partial$*; in a strict agent,* $\Delta = \mu AExts \circ \partial$*; in a conservative agent,* $\Delta = \nu AExts \circ \partial$*, where* ν *is minimization with respect to a quasi-order* $\preceq$ *on transitions whose minima include the null transitions; and in a strictly conservative agent,* $\Delta = \nu\mu AExts \circ \partial$*.*

(33.2) COROLLARY. *If* $AExts = \mu AExts$*, plain agents are strict, and conservative agents are strictly conservative.*

(33.3) QUESTION. *Are there interesting cases in which strictness and conservation interact, for example (1)* $S \subseteq S', S''$ *and* $(S, S') \preceq (S, S'')$ *implies* $S' \subseteq S''$*, or (2)* $S \subseteq S', S''$ *and* $S' \subseteq S''$ *implies* $(S, S') \preceq (S, S'')$*? What can be concluded about monotone kernel agents, that is, agents in which* $\partial(S_t, e_t) \subseteq \partial(S_{t+1}, e_{t+1})$ *at every successive step (as suggested in the monotone embedding theorem)?*

The question of computational complexity is particularly vexing, for while much work has been done on practical systems that seem to revise simple reasons states in acceptable times, the precise actions and costs of these algorithms are not yet known.[18] The principal practical difficulty to be investigated is whether the two sorts of minimization can be efficiently mechanized separately and in combination. RMS and its relatives operate as finitely grounded, hence strict, agents, and they attempt to approximate conservation. On the other hand, the strictness of RMS was in part a reaction to earlier, more purely conservative systems which did not observe the prudence of strict inference needed in some axiomatic systems, so it may be that conservative agents are efficiently mechanizable as well. I suspect that some strictly conservative agents have acceptably efficient mechanizations, but I have no algorithms or proofs to offer.

In general, however, it appears that strict conservatism may be more difficult to realize than either strictness or conservatism separately. Let us again suppose the agent finite, that $(S_1, S_2) \preceq (S_3, S_4)$ is deterministically computable in time polynomial in the sizes of the states involved, as is the case for the symmetric difference comparison. We again assume and ignore a constant environment, and suppose that $\partial(S)$ is deterministically computable in time polynomial in the size of S.

(33.4) THEOREM. *If* $(\mathscr{S}, \Delta)$ *is a plain simple reasons agent, then Is* $E \in \Delta(S)$*? is in P.*

PROOF. First compute $\partial(S)$, and check, as before, $E \in QExts(\partial(S))$ or $E \in FGExts(\partial(S))$, depending on whether the agent is locally grounded or finitely grounded. Each of these steps is in P, so their combinations are also. ∎

(33.5) THEOREM. *If* $(\mathscr{S}, \Delta)$ *is a strict finitely grounded simple reasons agent, then Is* $E \in \Delta(S)$*? is in P.*

PROOF. Since $\mu FGExts = FGExts$, the previous theorem applies. ∎

(33.6) THEOREM. *If* $(\mathscr{S}, \Delta)$ *is a strict locally grounded simple reasons agent, then Is* $E \in \Delta(S)$*? is in co-NP.*

PROOF. This is just the question Is $E \in \mu QExts(\partial(S))$? Since ∂ is in P, by Theorem 18.65 these strict extensions are in co-NP. ∎

(33.7) THEOREM. *If* $(\mathscr{S}, \Delta)$ *is a conservative simple reasons agent, then Is* $E \in \Delta(S)$*? is in co-NP.*

PROOF. We see Is $E \notin \Delta(S)$? is in NP for locally (resp. finitely) grounded agents as follows. First compute $\partial(S)$, and accept if $E \notin QExts(\partial(S))$ (resp. $E \notin FGExts(\partial(S))$). If $E \in QExts(\partial(S))$ (resp. $E \in FGExts(\partial(S))$), pick $E' \in QExts(\partial(S))$ (resp. $E' \in FGExts(\partial(S))$) and accept if $(S, E') \preceq (S, E)$ but $(S, E) \not\preceq (S, E')$. ∎

[18] See [DOYLE 1979], [STALLMAN AND SUSSMAN 1977], [LONDON 1978], [MCALLESTER 1980], [CHARNIAK, RIESBECK, AND MCDERMOTT 1980], [THOMPSON 1979], [MARTINS 1983], [GOODWIN 1982], and [MCDERMOTT 1982B].

(33.8) THEOREM. *If $(\mathscr{S}, \Delta)$ is a strictly conservative finitely grounded simple reason agent, then Is $E \in \Delta(S)$? is in co-NP.*

PROOF. Since $\mu FGExts = FGExts$, the previous theorem applies. ∎

(33.9) THEOREM. *If $(\mathscr{S}, \Delta)$ is a strictly conservative locally grounded simple reasons agent, then Is $E \in \Delta(S)$? is in Π_2^p.*

PROOF. By putting the previous proofs together, we see that $E \notin \nu(S, \mu QExts(\partial(S)))$ is in NP given verifications, each co-NP, of $E \in \mu QExts(\partial(S))$. ∎

(33.10) CONJECTURE. *If $(\mathscr{S}, \Delta)$ is a conservative simple reasons agent, then Is $E \in \Delta(S)$? is NP-hard.*

(33.11) QUESTION. *Are there interesting conservation notions which are efficiently and incrementally computable? That is, if symmetric difference conservation means intractability, are there less stringent notions which admit good algorithms? Is the "standard reason enumeration" technique employed in RMS a good approximation to symmetric difference?*

Also deserving attention are special cases of these agents and approximate algorithms. Most practical systems, for example, base their claims of efficiency by requiring that only coherent sets are manipulated. I have grown increasingly suspicious of such restrictions of attention, but the question deserves proper treatment. Even if the complexity of strictly conservative agents is demonstrably intractable, there may still be approximate algorithms of reasonable efficiency. Since the whole practice of reasoned assumption-making is based on correcting errors, we can afford in practice to be tolerant of occasional imprudences in jumping to conclusions. Unfortunately, one cannot hope to measure degrees of approximate correctness in terms of the number of poor assumptions in a state, since just one mistakenly included assumption can have arbitrarily many uncontroversial but mistaken consequences. Instead, one must look to the relative frequencies of correctly to incorrectly computed extensions among the entire set of such computations, that is, one must treat the question of approximation as a question about probabilistic algorithms.

Not only have complexity questions remained unstudied, but even the appropriate complexity measures still require proper formulation. The usual measures of time and space on a RASP are of course important, but so are some specific natural measures. One can ask how many state components are reconsidered or examined in the transition process. This is simply expressed in terms of the sum of the sizes of the mention sets of all elements against which the state is checked. When reasons are represented as graph structures, as in existing mechanizations, this measure corresponds to the number of edges traversed by the algorithm. In addition to average and worst-case measures, the complexity relative to the number of elements changed also holds interest. Folklore has it that any algorithm must be arbitrarily bad with respect to this measure, in that one can choose transitions requiring unbounded numbers of element reappraisals, but which lead to only a bounded number of changes.

§**34.** We now turn from individual transitions to the global perspective of trajectory spaces. The *trajectories* of an agent are finite or infinite sequences of admissible states that observe the transition table; formally, sequences $\langle S_i \rangle_{i=0}^{n}$ ($n \leq \omega$) such that $S_i \in \mathscr{S}$ if $0 \leq i \leq n$, and if $i + 1 \leq n$, then $S_{i+1} \in \Delta(S_i, e)$ for some $e \in \mathcal{E}$. The trajectory space $\mathcal{T}(\mathscr{S}, \mathcal{E}, \Delta)$ is the set of all trajectories of the agent $(\mathscr{S}, \mathcal{E}, \Delta)$. This for simplicity assumes the world obeys no laws other than the agent's transition table. In the general case, we must consider the agent as part of an isolated system, with the world having admissible states $\mathscr{S}_W \subseteq \mathscr{S} \times \mathcal{E}$, and a given trajectory space $\mathcal{T}_W \subseteq (\mathscr{S}_W^*)$. In this case, the agent's trajectory space will be the projection of $\mathcal{T}_W$ onto $\mathscr{S}^*$, onto a subset of $\mathcal{T}(\mathscr{S}, \mathcal{E}, \Delta)$. We employ the notation $T; T'$ to mean concatenation of two trajectories, so that if $T = \langle S_0, \ldots, S_m \rangle$ and $T' = \langle S'_0, \ldots, S'_n \rangle$, then $T; T' = \langle S_0, \ldots, S_m, S'_0, \ldots, S'_n \rangle$. Of course, $\mathcal{T}$ need not be closed under concatenation. If $S \in \mathscr{S}$, we write $S; T$ and $T; S$ to mean, respectively, $\langle S \rangle; T$ and $T; \langle S \rangle$.

Many standard questions about dynamical systems arise here, such as reachability, the existence of trajectories connecting two states; the existence of limit sets or attractors to which all trajectories from some neighborhood

eventually lead; the existence of cycles (closed trajectories); the existence of limit cycles; stability of limits; the complexity of computations; and more. At the moment, results are lacking concerning most of these questions, so we cannot treat them further here. But these questions call out for study, since some of them are intimately tied up with familiar psychological questions. For example, if one's trajectories are defined by some sort of learning or searching procedure, then the existence and reachability of limit sets corresponds to the learnability of certain concepts or skills, or to the solvability of certain goals. This is all familiar from popular treatments of hill-climbing, where the learning or searching procedure is a gradient vector field on the state space. Furthermore, the complexity or length of trajectories to these limits might shed light on power laws or laws of diminishing returns, since diminishing returns indicate the existence of limit states unreachable by finite trajectories, and power laws are just particular shapes for curves of diminishing returns. The structure of trajectory space $\mathcal{T}(\mathscr{S}, \mathcal{E}, \Delta)$ is also closely connected with the structure of the state space $\mathscr{S}$, since many of the questions previously addressed about assumability and realizability can be cast directly as questions about reachability by monotone kernel trajectories. Trajectory space also serves as a model for various modal or temporal logics concerned with the evolution of properties of elements or states (arguability, coherence, etc.). We must forgo discussion of most of these topics. In the following, we treat only two of the topics connected with the global viewpoint: the connections between catastrophes and backtracking, and the construction of "subjective probabilities" from trajectory space.

§**35.** The preceding development has focussed on reasoned changes of state, in which the agent decides what shape its next state should take, specifies those qualities via ∂, and then solves the specifications in one way or other to move to a new state in Δ. But unless we ensure that all transition specifications have admissible solutions, it can happen that there are no potential successors in Δ. The incoherence of $\partial(S, e)$ may be accidental, or it may be deliberate. For example, if ∂ includes a contradiction in the specifications for the next state after S, it ensures $\Delta(S, e) = \emptyset$. In this use, the contradiction acts as an admission that "I can't go on like this!" But how should we take such accidents and admissions? As natural death and suicide? While those are possibilities, another is to take these incompetences of Δ as occasions for other sorts of state changes. We treat the former possibilities elsewhere, and pursue the latter here. We add to our store of constitutive assumptions for characterizing agents by introducing the *extended transition table* $\mathbf{E} : \mathcal{T}(\mathscr{S}, \mathcal{E}, \Delta) \times \mathcal{E} \to \mathbf{P}\mathscr{S}$, which we require to agree with Δ whenever Δ is nonsingular, that is, for every trajectory $T \in \mathcal{T}(\mathscr{S}, \mathcal{E}, \Delta)$, if $T = T'; S$ and $\Delta(S, e) \neq \emptyset$, then $\mathbf{E}(T, e) = \Delta(S, e)$. With this new constitutive assumption, agents are characterized by choices of $(\mathcal{D}, \mathcal{I}, \mathscr{S}, \mathcal{J}, \lhd, \mathcal{E}, \partial, \Delta, \mathbf{E})$.

The most trivial choices for $\mathbf{E}$ are $\mathbf{E}(T; S, e) = \Delta(S, e)$, in which case $\mathcal{T}(\mathscr{S}, \mathcal{E}, \mathbf{E}) = \mathcal{T}(\mathscr{S}, \mathcal{E}, \Delta)$ and there is not much to say, and $\mathbf{E}(T; S, e) = \mathscr{S}$ if $\Delta(S, e) = \emptyset$, in which case the agent executes a random walk through trajectory space when it is not following Δ. Some more interesting choices for $\mathbf{E}$ involve the traditional notion of backtracking. Backtracking originated as a means for outwitting pursuers. One travelled leaving an obvious trail until a stream or other natural trail-obscuring obstacle was reached. One then carefully walked backwards over the path just travelled, placing feet in the footsteps just left, until one reached a point at which one could take off in a new direction, leaving the actual trail chosen to be unobvious when compared with the false trail. This technique was adopted in artificial intelligence to elude the hounds of failure, by keeping track of the choice points encountered during a search of a space, and upon failure of one search path, resuming the search with one of the alternate paths indicated in the choice points. We reproduce this idea formally as follows, where we for simplicity assume an isolated agent. Let $T = \langle S_i \rangle_{i=0}^{n}$ ($n \leq \omega$) represent the trajectory of the agent, where S_i is the state at instant i. We define $Alt(T, i+1)$, the *alternatives* in T at instant $i+1$, by $Alt(T, i+1) = \Delta(S_i) - \{S_{i+1}\}$ if $\Delta(S_i) \neq \emptyset$. These then are the states that could have been pursued but were not. We get two backtracking regimes by assuming the agent falls back to any or the closest alternative state when necessary, that is, by defining $\mathbf{E}(T; S)$ to be $\Delta(S)$ whenever $\Delta(S) \neq \emptyset$, and to be $Alt(T, k)$ or $\nu Alt(T, k)$ otherwise, where $k = \max\{j < i \mid Alt(T, j) \neq \emptyset\}$. Call these *chronological backtracking* and *conservative chronological backtracking*. Conservative non-chronological backtracking (also known as "dependency directed" backtracking) arises by taking $\mathbf{E}(T; S)$ at singular points to be $\nu \bigcup_{j<i} Alt(T, j)$. These definitions are simpler than those commonly employed since they do not remove explored paths from the set of alternatives.

To return to the example of §32, suppose $S_4 = \partial(S_3) \cup B \cup C \cup F$. We then have $Alt(\langle S_0, S_1 \rangle) = Alt(\langle S_0, S_1, S_2 \rangle) = Alt(\langle S_0, S_1, S_2, S_3 \rangle) = \emptyset$ and $Alt(\langle S_0, S_1, S_2, S_3, S_4 \rangle) = \partial(S_3) \cup B \cup C \cup G$.

§**36.** One way of understanding the nature of backtracking is in terms of the notion of catastrophe. (The following remarks are more suggesive than substantive at present. I hope to justify them rigorously elsewhere.) For our purposes here, a catastrophe is a "discontinuous" jump in a trajectory. Each time Δ is multi-valued, we have the possibility of trajectory space branching into two "sheets." Actually, not all multi-valued occurrences of Δ need produce bifurcations in trajectory space if the alternatives are all reachable from each other by ordinary trajectories. But a "true" bifurcation is introduced if the resulting states are mutually unreachable in $\mathcal{T}(\mathscr{S}, \mathcal{E}, \Delta)$. If an ordinary trajectory comes up to the edge of one of these sheets, the only way to proceed is by means of a jump to a state on another sheet, a "discontinous" change in the sense that the ordinary transition table provides no direction, which one might interpret as lack of a "derivative" with which the agent might predict its possible actions. There are many ways of embedding the ordinary trajectory space in larger spaces that have no discontinuities, and we have seen a few of these in the extended transition table **E**. If the agent is conservative, it might jump to one of the closest points on the other sheets. Alternatively, it might jump to a point corresponding to an alternative future, a point at the same time but on a different sheet among those bypassed along the actual trajectory. One might call this "sidetracking" instead of "backtracking." In any event, rules for continuing all trajectories beyond singularities have the effect of pasting all the edges in trajectory space back into the space. This results in a complex shape for the enlarged trajectory space, just as pasting all the edges of a sheet of paper together in the right way produces a sphere, a torus, or a Klein bottle, surfaces with no edges.

The most important quality of a catastrophe in a psychology is that, from the agent's point of view, it "just happens." While we may design agents to suffer catastrophes in certain ways, the actions involving the agent's own deliberation and choice are all captured by ∂ and Δ, and changes not involving those are beyond its powers. Of course, an agent may deliberately place itself in a position where, it knows, a catastrophe is inevitable, so as to chance some action not normally doable, just as one might lead one's self into the depths of despair, thinking that either death or a jump to some redeeming faith are the only possible paths, in spite of calm inability to believe. This difference in the quality of these sorts of actions has never admitted clear articulation in artificial intelligence before. In my own work on RMS, I always thought that the two processes of reason maintenance and backtracking were different, but could not say why. MCALLESTER developed ways of mechanizing them with a uniform procedure, and I thought that wrong. Instead, I was wrong, in that however one best mechanizes the two processes, by two mechanisms or by one, the processes still are different as far as the agent is concerned. Before, I had no clear way of distinguishing the agent and its actions from its realization and the changes that happen to it. The distinction between these processes lies in the former realm, MCALLESTER'S unification in the latter.

§**37.** With exact characterizations of theories of reasoned assumptions and their revision, we can connect artificial intelligence treatments of uncertainty with standard probabilistic treatments. These connections divide into two parts, one static, one dynamic. Just as theories of reasoned assumptions divide into characterizations of the set of admissible states and characterization of the temporal evolution of the agent's state, probabilistic theories divide into characterizations of particular states as probability distributions and characterization of the effects of actions as the evolution of the agent's probability distribution. We treat both parts below.

Drawing theoretical connections between these two approaches to uncertainty is not simply an abstract mathematical exercise. Instead, these connections yield two valuable benefits. The first benefit is the possibility of justifying the claimed methodological superiority of the artificial intelligence treatment over standard probabilistic treatments. Several authors in artificial intelligence have explained their abandoning probability theory in terms of the unpleasantly large amounts of information and computation required to apply the probabilistic theories.[19] With exact connections between the formal theories, we can begin to justify (or refute) these intuitions rigorously with the methods of computational complexity. The second benefit is that probability theory, whatever its computational disadvantages, offers certain sorts of information often useful but not present in the preceding theories of reasoned assumptions. In particular, probability theory allows one to summarize one's uncertainty about some question with a simple description, a number, and to compare degrees of uncertainty by comparing these numbers. The exact connection between the two theories permits us to recover degrees of certainty when desired even if we base the structure and action of systems on reasoned assumptions.

§**38.** The basic idea underlying the connection between reasoned assumptions and degrees of certainty is to make the degree of certainty of some state component e with respect to a set S of reasons depend on how e appears in admissible extensions of S. There is no distinguished way of defining this dependence, but instead a variety of possible measures. For example, one might ask how likely is e to occur in a randomly selected admissible extension, or how many arguments there are for e, or how many assumptions e depends on, etc. Each of these captures different intuitions about the meaning of "degree of certainty," and may be preferred in different circumstances.

These measures may be motivated in two ways. In the first, we view the nondeterministic theories of reasoned assumptions presented earlier as incomplete specifications of deterministic agents, whose indeterminism we interpret probabilistically in terms of how likely our incomplete specifications are to predict the actual behavior among the predicted possible behaviors. Alternatively, we can treat these measures as specifications for probabilistic algorithms.[20] The problem of acting prudently is too difficult or even impossible to solve exactly, so the agent makes a series of random choices (of admissible assumptions rather than numbers) and then decides what to do by exact means. These random choices may lead to error, so the task of the agent is to constrain the range of possible sets of choices (by using ratiocinative rules of thumb) so that the expected probability of success is high. This latter view seems well-suited to the mechanization of artificial agents.

§**39.** Perhaps the simplest way of measuring how the agent holds some state component is to assign a weight to each state the agent might be in, a weight representing how likely the agent is to be in that state, and then to sum the weights of all states containing the element. That is, if $w(S)$ is the weight of state S, then the measure of how e occurs in $\hat{S} \subseteq \mathscr{S}$ is just

$$\sum_{S \in \hat{S}} w(S).$$

[19] See [SZOLOVITS 1978], [DOYLE 1983B], and the discussion below.
[20] [RABIN 1976]

Weight functions like this are ordinarily expressed as *measure functions* on sets of states, as non-negative, additive functions $m : \mathbf{P}\mathscr{S} \to \mathbf{R}$, where $w(S) = m(\{S\})$. Nonnegative, as usual means that $m(\hat{S}) \geq 0$ for every $\hat{S} \subseteq \mathscr{S}$, and additive means that for every $\hat{S}, \hat{S}' \subseteq \mathscr{S}$, we have $m(\hat{S} \cup \hat{S}') = m(\hat{S}) + m(\hat{S}') - m(\hat{S} \cap \hat{S}')$. Actually, technical complications make this definition adequate only for the case of finite $\mathscr{S}$, but that is enough for most of our purposes. Later we comment on interesting cases of infinite $\mathscr{S}$.

For example, one natural measure function is the *counting* measure $m^\dagger$ that gives every state equal weight, i.e. $m^\dagger(\hat{S}) = |\hat{S}|$. This measure corresponds to making LAPLACE'S assumption, that every possible state of the agent is equally likely to occur. With this measure, the degree of certainty of a state component in a set of admissible extensions is just the percentage of them in which it appears.

Another natural measure function is the *specificity* measure m^* that weights states by how "specific" they are, namely $m^*(\{S\}) = 2^{-|S|}$. One way of looking at this measure is to think of states as partial descriptions of all the sets of components extending them, and to weight states proportionally to the number of possible supersets. From this point of view, we would define $m^{*\prime}(\{S\}) = 2^{|\mathcal{D}|-|S|}$. This, of course, simply multiplies all of the specificity weights by the constant $2^{|\mathcal{D}|}$, and so changes none of the comparative relations between weights of states.

Yet another measure function is the *selection* measure $m^!$ that weights states as selections of elements from the domain, namely $m^!(\{S\}) = \binom{|\mathcal{D}|}{|S|}^{-1}$. We make no important use of this measure in the following.

§**40.** For each theory of reasoned assumptions and measure function m, we define the "degree of certainty."

(40.1) DEFINITION. *Let $\hat{S} \setminus A = \{S \in \hat{S} \mid A \subseteq S\}$ for $\hat{S} \subseteq \mathscr{S}$. Then the extent $\mathcal{E}(A, S)$ of a set $A \subseteq \mathcal{D}$ relative to a set $S \subseteq \mathcal{D}$ is given by*

$$\mathcal{E}(A, S) = m(AExts(S) \setminus A)/m(AExts(S)),$$

where $\mathcal{E}(A, S) = 1$ whenever $AExts(S) = 0$.

Thus the extent of a state component e is just the relative measure $\mathcal{E}(\{e\}, S)$ of those admissible extensions containing e.

(40.2) COROLLARY. *If $A \subseteq \mathcal{D}$ is not arguable in S, then $\mathcal{E}(A, S) = 0$, and if A is inevitable in S, then $\mathcal{E}(A, S) = 1$.*

(40.3) DEFINITION. *If S is coherent, $d, e \in \mathcal{D}$, and $\mathcal{E}(\{d\}, S) > 0$, then $\mathcal{CE}(e \mid d, S)$, the conditional extent of e given d in S, is*

$$\begin{aligned} \mathcal{CE}(e \mid d, S) &= \mathcal{E}(\{d, e\}, S)/\mathcal{E}(\{d\}, S) \\ &= m(AExts(S) \setminus \{d, e\})/m(AExts(S) \setminus \{d\}). \end{aligned}$$

I am not sure what approach to take for incoherent S. If S is coherent but $\mathcal{E}(\{d\}, S) = 0$, the situation here has analogies to the situation in Bayesian probability theory that motivates so-called *Popper functions*, but I leave treatment of that case to future studies as well.

(40.4) THEOREM (BAYES). *If $\{d_i\}_{i=0}^n$ is a subset of $\mathcal{D}$ such that exactly one d_i occurs in every admissible extension of S, then*

$$\mathcal{CE}(d_i \mid e, S) = \frac{\mathcal{CE}(e \mid d_i, S) \cdot \mathcal{E}(\{d_i\}, S)}{\sum_{j=0}^n \mathcal{CE}(e \mid d_j, S) \cdot \mathcal{E}(\{d_j\}, S)}.$$

(40.5) COROLLARY. *If at most one d_i occurs in every admissible extension of S, then*

$$\mathcal{CE}(d_i \mid e, S) \leq \frac{\mathcal{CE}(e \mid d_i, S) \cdot \mathcal{E}(\{d_i\}, S)}{\sum_{j=0}^{n} \mathcal{CE}(e \mid d_j, S) \cdot \mathcal{E}(\{d_j\}, S)}.$$

Linguistic reasons agents are of special interest here, for their formulation involves many of the same assumptions as does subjective Bayesian decision theory. Because state components have logical form in this theory, we can examine the relations between extents of components with related forms.

(40.6) THEOREM. *Let $S \subseteq \mathcal{D}$. Then*
(i) If S is not inevitably consistent, then for each $x \in \mathcal{D}$, $\mathcal{E}(x, S) = 1$,
(ii) S is inevitably consistent iff for each $x \in \mathcal{D}$, $\mathcal{E}(x, S) + \mathcal{E}(\neg x, S) \leq 1$,
(iii) For each $x, y \in \mathcal{D}$, $\mathcal{E}(x \wedge y, S) \leq \mathcal{E}(x, S) + \mathcal{E}(y, S)$,
(iv) For each $x, y \in \mathcal{D}$, if $x \supset y$ is inevitable, then $\mathcal{E}(x, S) \leq \mathcal{E}(y, S)$, and
(v) If $S \neq \{\mathcal{D}\}$, then for each $x, y \in \mathcal{D}$, $\mathcal{E}(x, S) \geq \mathcal{E}(x \wedge y, S) + \mathcal{E}(x \wedge \neg y, S)$.

PROOF. Let $S \subseteq \mathcal{D}$ and $\hat{S} = AExts(S)$. (i) If S is incoherent, then $\mathcal{E}(x, S) = 1$ by definition, and if S is inevitably inconsistent, $\hat{S} = \{\mathcal{D}\}$, so again $\mathcal{E}(x, S) = 1$ for every $x \in \mathcal{D}$. (ii) follows immediately from this last observation. (iii) Since $\{x \wedge y\}$ is interdeducible with $\{x, y\}$, $\hat{S} \setminus \{x \wedge y\} \subseteq \hat{S} \setminus \{x\} \cup \hat{S} \setminus \{y\}$. (iv) If $\hat{S} = \emptyset$, the claim follows trivially. If $\hat{S} \neq \emptyset$, then $\hat{S} \setminus \{x\} \subseteq \hat{S} \setminus \{y\}$ since $\{x, x \supset y\} \vdash \{y\}$. (v) If $\hat{S} \neq \{\mathcal{D}\}$ then each $S \in \hat{S}$ is consistent. Since $\{x, z\}$ and $\{x \wedge z\}$ are interdeducible, we must have $\hat{S} \setminus \{x, y\} \cap \hat{S} \setminus \{x, \neg y\} = \emptyset$. Since $\hat{S} \neq \emptyset$, the claim follows. ∎

Since arguability is different from assumability, we introduce a parallel notion to conditional extents, namely *a posteriori extents*.

(40.7) DEFINITION. *Let $d, e \in \mathcal{D}$ and $S \subseteq \mathcal{D}$. If $S \cup \{e\}$ is coherent, then $\mathcal{AE}(d \mid e, S)$, the a posteriori extent of d given e in $\hat{S}$, is*

$$\mathcal{AE}(d \mid e, S) = \mathcal{E}(\{d\}, S \cup \{e\}).$$

Unfortunately, I do not know how to treat the case of incoherent $S \cup \{e\}$. There also seems to be no general way to relate conditional and *a posteriori* extents. This weakness of the theory is to be expected from the complexity of the ways in which the sets of admissible extensions may change under the addition of new information. Stronger theories may be possible for restricted sorts of agents, but by and large these are unexplored.

§**41.** Let us examine some examples to see how these measures differ. We work within a simple reasons agent with $\mathcal{D} = \{c_1, \neg c_1, c_2, \neg c_2, r_1, r_2, r_3, r_4, r_5, r_6\}$, where $\mathcal{I}(c_1) = \mathcal{I}(\neg c_1) = \mathcal{I}(c_2) = \mathcal{I}(\neg c_2) = \mathbf{P}\,\mathcal{D}$, and where $\mathcal{I}(r_1) = \emptyset \setminus\!\setminus \{\neg c_1\} \mid\!\!\sim \{c_1\}$, $\mathcal{I}(r_2) = \emptyset \setminus\!\setminus \{c_1\} \mid\!\!\sim \{\neg c_1\}$, $\mathcal{I}(r_3) = \{c_1\} \setminus\!\setminus \{\neg c_2\} \mid\!\!\sim \{c_2\}$, $\mathcal{I}(r_4) = \{c_1\} \setminus\!\setminus \{c_2\} \mid\!\!\sim \{\neg c_2\}$, $\mathcal{I}(r_5) = \emptyset \setminus\!\setminus \{\neg c_2\} \mid\!\!\sim \{c_2\}$, and $\mathcal{I}(r_6) = \emptyset \setminus\!\setminus \{c_2\} \mid\!\!\sim \{\neg c_2\}$.

Let $S = \{r_1, r_2, r_3, r_4\}$ and $S' = \{r_1, r_2, r_5, r_6\}$. Then S has three admissible extensions

$$\begin{aligned} E_1 &= S \cup \{\neg c_1\}, \\ E_2 &= S \cup \{c_1, c_2\}, \\ E_3 &= S \cup \{c_1, \neg c_2\}, \end{aligned}$$

and S' has four admissible extensions

$$\begin{aligned} F_1 &= S' \cup \{c_1, c_2\}, \\ F_2 &= S' \cup \{c_1, \neg c_2\}, \\ F_3 &= S' \cup \{\neg c_1, c_2\}, \\ F_4 &= S' \cup \{\neg c_1, \neg c_2\}. \end{aligned}$$

We then have the following extents and conditional extents.

$\mathcal{E}(x, S)$	$m^\dagger$	m^*	$m^!$
c_1	$2/3$	$1/2$	$.70+$
$\neg c_1$	$1/3$	$1/2$	$.29+$
c_2	$1/3$	$1/4$	$.35+$
$\neg c_2$	$1/3$	$1/4$	$.35+$
r_1	1	1	1
r_5	0	0	0

$\mathcal{E}(x, S')$	$m^\dagger$	m^*	$m^!$
c_1	$1/2$	$1/2$	$1/2$
$\neg c_1$	$1/2$	$1/2$	$1/2$
c_2	$1/2$	$1/2$	$1/2$
$\neg c_2$	$1/2$	$1/2$	$1/2$
r_1	0	0	0
r_5	1	1	1

In the following, entries represent $(m^\dagger, m^*)$.

$\mathcal{CE}(x \mid y, S) =$

$x \backslash y$	c_1	$\neg c_1$	c_2	$\neg c_2$	r_1
c_1	$(1,1)$	$(0,0)$	$(1,1)$	$(1,1)$	$(2/3, 1/2)$
$\neg c_1$	$(0,0)$	$(1,1)$	$(0,0)$	$(0,0)$	$(1/3, 1/2)$
c_2	$(1/2, 1/2)$	$(0,0)$	$(1,1)$	$(0,0)$	$(1/3, 1/4)$
$\neg c_2$	$(1/2, 1/2)$	$(1,1)$	$(0,0)$	$(1,1)$	$(1/3, 1/4)$
r_1	$(1,1)$	$(1,1)$	$(1,1)$	$(1,1)$	$(1,1)$
r_5	$(0,0)$	$(0,0)$	$(0,0)$	$(0,0)$	$(0,0)$

Constructions similar to the above were first proposed by CARNAP in his theory of probability as "degrees of entailment."[21] While the ideas are similar, there are important differences between the two approaches. Let $\mathcal{D} = \{a, \neg a, b \neg b, r, \neg r\}$, $S = \{a, r\}$, and define $\mathcal{I}(a) = \mathcal{I}(\neg a) = \mathcal{I}(b) = \mathcal{I}(\neg b) = \mathbf{P}\,\mathcal{D}$, $\mathcal{I}(r) = \{a\} \setminus\setminus \{\neg r\} \mid\!\!\sim \{b\}$, and $\mathcal{I}(\neg r) = \emptyset \setminus\setminus \{b\} \mid\!\!\sim \{\neg b\}$. In this case, S has the sole admissible extension $\{a, r, b\}$, so all extents and conditional extents are either 1 or 0, no matter which measure function is chosen. But let us consider a posteriori

[21] [CARNAP 1950], [KYBURG 1970]

extents. Here we have the table, again independent of choice of measure function,

x	$\mathcal{E}(x,S)$	$\mathcal{AE}(x \mid y,S):$	a	$\neg a$	b	$\neg b$	r	$\neg r$
a	1		1	1	1	1	1	1
$\neg a$	0		0	1	0	0	0	0
b	1		1	1	1	1	1	0
$\neg b$	0		0	0	0	1	0	1
r	1		1	1	1	1	1	1
$\neg r$	0		0	0	0	0	0	1

Note here that while $\mathcal{E}(b,S) = 1$ and $\mathcal{E}(\neg b,S) = 0$, we have $\mathcal{AE}(b \mid \neg r,S) = 0$ and $\mathcal{AE}(\neg b \mid \neg r,S) = 1$, so accepting $\neg r$ involves "learning" $\neg b$ and "unlearning" b, no matter which of the measure functions we use. This ability to learn and unlearn things is quite different from the standard sorts of learning theories studied by CARNAP, and even in common cases, the differences in what objects are measured permits "inductive learning" with any of our measure functions, something not true of CARNAP'S theories. Unfortunately, we cannot pursue this topic here.

§**42.** One extension to the above constructions is to allow measures of how a state component appears in the admissible extensions of a set of reasons to depend on the element as well as on the admissible extensions. For example, if we want to weight components by how many arguments warrant their presence in a state, we cannot use a pure measure on states, for that would treat every element in the same way. To permit the wider class of measures of "certainty," we allow measure functions m_A dependent on subsets A of $\mathcal{D}$, subject to the additivity requirement that if $A, B \subseteq \mathcal{D}$ and $\hat{S} \subseteq \mathscr{S}$, then

$$m_{A\cup B}(\hat{S}) = m_A(\hat{S}) + m_B(\hat{S}) - m_{A\cap B}(\hat{S}).$$

In effect, we replace measure functions on $\mathbf{P}\mathscr{S}$ by measure functions on $\mathbf{P}\mathcal{D} \times \mathbf{P}\mathscr{S}$. Thus we might define a simplistic argument counting measure by $m_{\{e\}}(\{S\}) = |\{d \in S \mid d \text{ is a valid reason for } e \text{ in } S\}|$.

§**43.** With probabilistic constructions on sets of reasons, we can recast the description of the agent's evolution in time in terms of these constructions. To make the new constructions easier to grasp, we first treat the ahistorical, isolated system $(\mathscr{S}, \Delta)$ and then extend the construction to history-dependent, non-isolated systems $(\mathscr{S}, \mathcal{E}, \mathbf{E})$.

The basic idea of the probabilistic treatment of an agent's state-evolution is to view the "nondeterministic" transition table Δ instead as an incomplete specification of a probabilistic transition table. For example, under a Laplacian assumption, we could take each transition $S' \in \Delta(S)$ to be equally likely among the possible transitions from S. Alternatively, we could assume transitions to simple states are more likely than transitions to more complex states. In fact, each measure function m on $\mathbf{P}\mathscr{S}$ such that $m(\mathscr{S}) > 0$ gives rise to a probabilistic interpretation of Δ by defining the probability of moving from $S \in \mathscr{S}$ to $S' \in \Delta(S)$ to be $m(\{S'\})/m(\Delta(S))$. (For the time being, we assume that $\Delta(S)$ is nonempty for every $S \in \mathscr{S}$.) This interpretation recalls the relation between the standard and many-worlds interpretations of quantum mechanics, but we cannot pursue that here.[22]

With assumptions about transition probabilities captured in measure functions, we can describe the temporal evolution of an agent $(\mathscr{S}, \Delta)$ in terms of the probability that the agent will be in state S at time t. We assume initial probabilities of being in a state for time $t = 0$, and then apply the transition probabilities to determine the probabilities for times $t > 0$. The instantaneous functions assigning probabilities to states are called *density functions*, and are simply measure functions f on $\mathbf{P}\mathscr{S}$ such that $f(\mathscr{S}) = 1$, that is, functions that say that the agent must be in *some* state. As above, every nontrivial measure function m such that $m(\mathscr{S}) > 0$ induces a density function by defining $f(X) = m(X)/m(\mathscr{S})$. Thus beginning with an "initial state function," for example

$$f_0(X) = \begin{cases} 0 & S_0 \notin X \\ 1 & S_0 \in X \end{cases},$$

[22] [DE WITT AND GRAHAM 1973], [VAN FRASSEN 1980]

we look to calculate the sequence of density functions $\langle f_0, f_1, f_2, \ldots \rangle$ describing the evolution of the system $(\mathscr{S}, \Delta)$ with respect to m.

Note first that the set of all density functions on $\mathscr{S}$ is a convex set. That is, if f_1 and f_2 are density functions, and if $w_1 + w_2 = 1$, then $w_1 f_1 + w_2 f_2$ is a density function as well.

Successive density functions for a system are computed by combining the probability that the system is in a particular state with the probability of moving from that state to specific other states. We add up the probabilities of reaching a state, and we have the new density function. Formally, we define f_{t+1} for each $S \in \mathscr{S}$ by

$$f_{t+1}(\{S\}) = \sum_{x \in \mathscr{S}} f_t(\{x\}) \cdot \frac{m(\Delta(x) \cap \{S\})}{m(\Delta(x))}.$$

The resulting function f_{t+1} is a density function since by induction $\sum_{x \in \mathscr{S}} f_t(\{x\}) = 1$ and f_{t+1} is the convex combination of density functions $g_x(\{S\}) = m(\Delta(x) \cap \{S\})/m(\Delta(x))$.

As in the instantaneous case considered previously, we can extract "degrees of belief" from the successive density functions by defining, for $A \subseteq \mathcal{D}$, $c_t(A) = f_t(\mathscr{S} \setminus A)$ as the probability that the system is in a state containing the components A. As before, these projections of state probabilities onto components need not be unitary, since in a linguistic reasons agent we might have $c_t(\{d\}) + c_t(\{\neg d\}) < 1$.

We extend this construction to the case of nonisolated, possibly singular, extended transition tables as follows. Let $(\mathscr{S}, \mathcal{E}, \mathbf{E})$ describe the agent. In the general case, transition probabilities may depend upon the full history of the agent, and we represent some of these with a measure function F on $\mathcal{T} \times \mathcal{E} \times \mathcal{T}$, interpreting $F(\{T\}, \{e\}, \{T'\})$ as the probability of moving to trajectory T' given the previous trajectory T and environment e. Ideally, we would get F by projecting probability measures on the world's transitions in $(\mathscr{S}_W)^* \times (\mathscr{S}_W)^*$, but we do not treat that most general case here. We assume F is normalized so that

$$\sum_{S \in \mathbf{E}(T,e)} F(\{T\}, \{e\}, T; S) = 1$$

for every $T \in \mathcal{T}$ and $e \in \mathcal{E}$, thus making the probabilistic interpretation possible. We construct a sequence of density functions f_t on $\mathcal{E}^* \times \mathcal{T}$ from the transition probabilities so that $f_t(\hat{e}, T)$ represents the probability of the agent having traversed trajectory T in response to the sequence of environments $\hat{e} \in \mathcal{E}^*$. The initial density function f_0 is given, and we define f_{t+1} so that

$$f_{t+1}(\hat{e}; e, T) = \sum_{T' \in \mathcal{T}} f_t(\hat{e}, T') \cdot F(T', e, T).$$

These density functions on $\mathcal{T}$ can be projected onto $\mathscr{S}$ to give the probabilities of the agent being in specific states by defining $\phi_t(\hat{e}, S)$ for every $t, \hat{e}, S$ to be

$$\phi_t(\hat{e}, S) = \sum_{T; S \in \mathcal{T}} f_t(\hat{e}, T; S)$$

.

§**44.** Many researchers in artificial intelligence have abandoned the direct use of probabilistic representations on the basis of several intuitions. Briefly, they are (1) One usually has only fragments of information about a few questions. Coming up with a complete matrix of conditional probabilities for all questions is not feasible, while incrementally formulating rules of thumb for generalities, exceptions, etc., is quite feasible. Much of the work in "knowledge acquisition" focusses on elicting and subsequently editing rules of thumb, though sometimes with "certainty factors" whose close values are recognized to be meaningless. (2) One usually has only a few known

goals or desires, augmented occasionally by problem-oriented reasoning, rather than a complete utility function. (3) It is easier to work by adding and removing individual statements from a database or subdatabase than to continually recompute probability distributions. (4) When errors of information are revealed, or when new sorts of events are formulated, one must completely revise one's system of conditional probabilities. But these revisions must be based on some qualitative considerations, to which the numerical probabilities are merely fit. (See [DOYLE 1983B] for another discussion of these.) These intuitions can be summarized by saying that the usual Bayesian approach, while a fine system for formulations of decisions *post hoc*, or for mechanization of thoroughly explored, stable, narrow decisions, is both informationally and computationally infeasible in broad, changing systems. Bayesianism simply says nothing about the problems of incrementally formulating systems of probabilities, which are the main practical difficulties facilitated by the artificial intelligence approach. If these prejudices are justified, then the constructions presented previously offer the consolation that even if one abandons probabilities in direct use, one can always recover them when necessary, that artificial intelligence practice is not contradicting the Bayesians so much as being forced to work with less.

These intuitions would likely profit from formal exploration. For instance, Bayesians are familiar with the notion of qualitative probabilities. Indeed, the foundations of statistics develops axioms for these qualitative probabilities, and one shows that any consistent set of qualitative probabilities can be fit with a compatible system of numerical probabilities. One proves this fit unique given assumptions about the fineness and topological completeness of the qualitiative probabilities. The artificial intelligence approach in effect concentrates on these qualitative probabilities, abandoning *a priori* fineness and topological completeness assumptions.

More concretely, if we consider linguistic reason agents, we see the primary difference between the theory of subjective probability and the theory of extents turns on the question of completeness. For Bayesians, subjective probabilities are unitary, that is, the subjective probabilities of x and $\neg x$ sum to 1 for each x. In contrast, in the present theory some admissible extensions of S may contain neither x nor $\neg x$, so that $\mathcal{E}(x, S) + \mathcal{E}(\neg x, S) < 1$. The theories could be brought into agreement if we required completeness of extensions, that is, added to $\mathcal{R}$ the requirement that for each $d \in \mathcal{D}$ and $S \in \mathscr{S}$, either $d \in S$ or $\neg d \in S$. But this is a very peculiar requirement, for two reasons. First, the whole point of reasoned assumptions is to be able to complete one's set of beliefs with respect to some question when necessary, taking lack of information and incompleteness as normal. Requiring completeness of extensions is not in direct conflict with this motivation, but it does mean replacing every ordinary extension with the set of all possible completions of the extension. Second, it seems unwise to attempt to rule out paradoxical sentences at the outset. We can think about and phrase the Liar paradox, and some reasonable artificial agents should share our ability. But the character of paradoxical sentences is that neither they nor their negations may be part of a consistent theory, so if we want interesting languages of thought, we cannot accept the Bayesian requirement of completeness.

Bayesianism stems from important motivations, but overidealizes. While the above constructions indicate the naturalness and importance of the idea of strength of beliefs and other attitudes, it also casts doubts upon the Bayesian identification of degree of belief with subjective probability. The measure of degree of belief via extents is a perfectly good probability measure, but the projection of this measure onto the logical structure of states is not in general a probability measure without the specious axiom of completeness. As in quantum theory, the projected measure represents a lattice of possible events, and only represents a boolean lattice in the special case of complete states.[23]

Of course, the Bayesian might respond that our constructions do not capture subjective probabilities, but simply *lower bounds* on subjective probabilities, as in the DEMPSTER/SHAFER theory of evidence.[24] While this reply does not address the difficulties posed by paradoxical sentences, it may offer a way of reconciling the views, but we cannot pursue that here.[25] Instead, we observe that whatever the attractions of the stronger Bayesian theory for more competent agents, in computationally realized agents, the "lower bounds" offered by our construction may be the only reasonable choices for "degrees of belief." We see this in two ways. The first is that extents arise naturally as the end result of the operation of probabilistic algorithms, and in this way are not so much computational approximations to ideal quantities as objective events observed by the agent (even though the events are part of the

[23] [BIRKHOFF 1967], [BELTRAMETTI AND CASSINELLI 1981]
[24] [SHAFER 1976]
[25] See [LEWIS 1980].

agent itself). The second point is that the dynamical form of extents, densities, arise naturally in consideration of revision of state, for example in backtracking. At singularities in a trajectory continued by means of plain chronological backtracking, the possible successors are just the successors of the previous state not taken. That is, if $T = T'; S$ and $\Delta(S) = \emptyset$, then $\mathbf{E}(T) = \mathbf{E}(T') - \{S\}$. In this case, densities in $\mathbf{E}(T)$ are approximately extents in the kernel of $\mathbf{E}(T')$, and correspond to the resilience of state components, to the relative ease with which they may be avoided in successor states. If some density is large, most successors will contain the component in question, so it is difficult to avoid. Indeed, we can adopt this view at every point of a trajectory, singular or not, by considering imaginary discontinuities, by asking how the agent would have to view things were the current view forbidden. In this way, densities in $\mathbf{E}(T)$ can always be taken as the resilience of state components. This interpretation of densities is particularly appealing since it applies to all components of mental states uniformly. For example, in attitudinal theories it provides measures not only of degrees of belief but of strength of desire and firmness of intent. Pursuit of this interpretation leads to an interesting non-Bayesian decision theory, but that is beyond the scope of this paper.

If consideration of extents in trajectories provides motivations for some of the concerns of subjective Bayesian probability theory, it also helps understand ZADEH'S notion of fuzzy sets and concepts.[26] Statements like "Sue is tall" are considered vague because "tall" is not a well-defined concept; there are many heights Sue could reach and be thought tall. ZADEH formalizes this notion by introducing a spectrum of truth-values for the sentence, a spectrum derived from a spectrum of tallness-values. One might instead develop a theory of fuzzy concepts in terms of extents. Rather than simply *assuming* tallness spectra, one could formulate exact theories of tallness and simply look to see what distribution these entail for particular statements. For example, one might require exact theories of tallness to specify exact intervals of height, and given this restriction look to the admissible extensions of the statement "Sue is tall." If there are many intervals saying one height is tall and fewer intervals saying another height is tall, then the first height will be "more tall" than the second. Example 24.11 can be read as a formal example of this idea in terms of the concept "is a lot." In that example, 0 is definitely not a lot (lot to degree 0), each $n \geq 3$ is definitely a lot (lot to degree 1), while 1 and 2 have intermediate values: 2 is twice as much a lot as is 1 (lot to degree $\frac{2}{3}$ versus degree $\frac{1}{3}$).

If one approach is to have advantages over the other in practice, the advantages may depend far more on practical utility than on theoretical elegance. The two most important questions are how difficult it is to obtain the information required by each approach, and how costly are the computations involved. As [DOYLE 1983B] mentions, the Bayesian informational requirements are severe because one needs vast amounts of information, few bits of which are easily had from experts. The needed information may be had much more easily in the form of ratiocinative rules of thumb. But complementing this, the probabilistic constructions facilitate the use of information acquired as rules of thumb, since they permit comparison of relative strengths or certainties of state components. Since these measures reflect the overall structure of the set of admissible extensions of some kernel, they may be useful in summarizing that structure in lieu of analytical characterization of its branchings, alternatives, etc. Unfortunately, practical exploitation of these possibilities demands efficient algorithms for computing admissible extensions according to specified measure functions. At present, algorithms for this purpose are completely unstudied. Systems like RMS make arbitrary choices in constructing admissible extensions, and while several authors have suggested making these choices depend on properties of the forseen resulting extensions, there is no known way of deriving the resulting measure on states from these intra-algorithmic choices, nor any known recipe for finding an appropriate algorithm when the final measure is specified. The mix of techniques used in practice will depend on the relative facilities and computabilities of the Bayesian and reasoned assumptions approaches. Reasoned assumptions facilitate some important operations, and Bayesianism facilitates others. Can one relate the total cost of working within one approach to the total cost of the other?

[26] [ZADEH 1975]

IV. Related Theories of Reasoned Assumptions

§**45.** The previous chapter included both development of a theoretical framework for describing theories of reasoned assumptions as well as some first steps toward a mathematical development of the framework itself. The value of such a framework, if any, must begin with the power to clearly describe specific theories, rather than with the possibility of a mathematical theory based on the framework, although once the grounds of the theory are secured, mathematical analysis can reveal hidden structure. We have seen how the framework eased the introduction of the several constitutive ideas presented previously. In the following sections we further exercise these concepts by using them to analyze and summarize a variety of systems in artificial intelligence touching on notions related to reasoned assumptions. Unfortunately, since several of these systems have never been exactly formulated or described in the literature, or have been described only in terms of the behavior of a set of complex procedures, we cannot always rigorously justify our analyses. Instead, one of the benefits of the current approach is that it may allow authors to precisely specify the intended structure and behavior of their artificial intelligence systems, whether before, during, or after development of the systems.

§**46.** The first analysis concerns RMS, a program developed by the present author, but standing among several related programs in historical order.[27] The purpose of RMS ("reason maintenance system;" originally in the literature as TMS, "truth maintenance system") is to record and revise a set of database entries, carrying out these activities at the behest of a substantive program. RMS performs these duties by recording and analyzing reasons or justifications for database entries in terms of other entries.

As a good first approximation, we can characterize RMS in terms of finitely grounded simple reasons and contradictions, strict transitions, an approximation to symmetric difference conservatism, and non-chronological backtracking. We justify these in turn as best we can. See [DOYLE 1983C] for a short, self-contained characterization of RMS.

The reasons or justifications of RMS are exactly of the form of finite simple reasons with singleton conclusions, that is, $A \setminus\setminus B \mid\!\!- \{c\}$. In RMS, the first set of elements is called the *inlist*, the second the *outlist*, and the sole element of the third the *conclusion*. In RMS, as in the simple reasons theory, the nature of the domain from which state components are drawn is left unspecified. RMS assigns a name to each domain component as it is presented to the system, and subsequently operates solely in terms of the internal name. Usually, these internal names are state components with trivial interpretations, but some of these may instead be designated contradictions with the empty interpretation. The only other state components are the reasons proper, which have the usual interpretation.

RMS constructs database states to be finitely grounded extensions of the current set of reasons exactly as in the simple reasons theory.

RMS revises the current state to assimilate new reasons added to the current set of reasons. It does this by temporarily removing from the state all components which might be affected by the new reasons, and by then adding back in as many as possible when valid reasons can be found among those left for components in limbo. Unfortunately, I know not how to exactly characterize the actual performance of RMS. It clearly realizes strict transitions because of the finite groundedness ensured by the remove-and-restore algorithm. The intent of the program also appears to be to construct a minimal perturbation with respect to symmetric difference conservatism. The principal predecessor of RMS would completely rederive the entire state following every subtractive kernel modification.[28] To avoid this, its successors employed "incremental" recomputation of just that state subset directly affected by the changes. Unfortunately, this does not mean RMS actually realizes symmetric difference conservatism, since its aim was never mathematically formulated, and its procedure makes on-the-fly choices that were hoped to approximate the above relations. Its approximations may in fact be good, since it always examines reasons for a particular conclusion in the same order, and so may reconstruct a previous extension subset if it had not really been affected by

[27] [DOYLE 1979], [DOYLE 1980]

[28] [STALLMAN AND SUSSMAN 1977]

the perturbation, but the procedures are complex enough to make this difficult to verify, and I bet there are counter-examples. It may be possible to correctly minimize the symmetric difference by recording the previous statuses of state components and attempting to reproduce them whenever choosing some partial revision, but that remains to be explored. RMS is an excellent example of a program suffering from ill-articulated and possibly ill-realized aims, in spite of conscientious care and labor in its design.

RMS employs non-chronological backtracking to avoid pursuing any state containing a contradiction. RMS backtracks by finding an alternative extension of its current state; precisely, by finding a likely perturbation and letting the ordinary revision procedures assimilate the perturbation. Unfortunately, here too I have no simple exact characterization of the process actually realized by RMS. Moreover, contradictions are not absolute, in that if RMS cannot find any alternative extension without the contradiction element (specifically, if it cannot be removed by additions to the current set of reasons), RMS simply proceeds in a state still containing the contradiction. Another difference is that while RMS recognizes and treats incoherence due to contradictions, RMS blindly fails to operate when the set of reasons is otherwise incoherent.

RMS also interprets some of its records as "conditional-proof justifications." These are counterfactual justifications, as in "I believe P because I would derive a contradiction were I to believe $\neg P$." These justifications can be formalized in terms of comparative similarity relations on states. We interpret a conditional-proof justification $(A \setminus\setminus B \mid\!\!\vdash C) \mid\!\!\vdash D$ by saying that D must be in S if $C \subseteq S'$ for every $S' \in \mathscr{S}$ such that $A \subseteq S' \subseteq B^c$ and S' is as similar as possible to S under the chosen notion of similarity of states. Formalized in this way, conditional-proof justifications are "oracles," involving difficult computational problems that RMS avoided by means of a complex half-measure. In any event, there are very interesting connections to be explored between the comparative similarity relations one might use in interpreting conditional-proof justifications and the comparative similarity relations derived from conservation relations. Understanding these connections might aid the correct and efficient mechanization of conditional-proof justifications.

§**47.** While our characterization of RMS suffered in accuracy due to the program's inarticulate development, our next application is much clearer. This is the logic of default reasoning as proposed by REITER, which turns out to be related to the linguistic reasons theory.[29] In the logic of default reasoning, states are composed of two sorts of elements: logical formulas, and defaults. Formally, $\mathcal{D} = \mathcal{D}_1 \oplus \mathcal{D}_2$, where $\mathcal{D}_1$ is the set of closed wffs of a first-order logical language $\mathcal{L}$, $\mathcal{D}_2$ is another set, and $\vdash$ is the ordinary deducibility relation on $\mathcal{L}$. The elements of $\mathcal{D}_1$ are called statements, the elements of $\mathcal{D}_2$ defaults. All statements are trivially interpreted, and all defaults are interpreted as special sorts of simple reasons involving only statements, specifically, if $d \in \mathcal{D}_1$, $\mathcal{I}(d) = \mathbf{P}\,\mathcal{D}$, and if $d \in \mathcal{D}_2$, there are (possibly open) wffs $a, c \in \mathcal{L}$, $B \subseteq \mathcal{L}$, such that

$$\mathcal{I}(d) = \{a\} \setminus\setminus \neg B \mid\!\!\vdash \{c\} = \bigcap_{\sigma \in \Sigma} \{S \subseteq \mathcal{D} \mid \sigma a \in S \subseteq (\neg \sigma B)^c \supset \sigma c \in S\},$$

where σ ranges over the set Σ of all substitutions of closed terms for free variables. We write ${}_1S$ and ${}_2S$ to mean $S \cap \mathcal{D}_1$ and $S \cap \mathcal{D}_2$ respectively. The only general restriction on states is that their statements be deductively closed, that is,

$$\mathcal{R} = \{S \subseteq \mathcal{D} \mid {}_1S = \mathrm{Th}({}_1S)\}.$$

As before, $AExts = FGExts$ where for every $d \in \mathcal{D}$ and $S \subseteq \mathcal{D}$,

$$\begin{aligned} \mathcal{J}(d, S) = \{E \mid d \in E \quad \supset \quad & [d \in S \ \vee\ E - \{d\} \vdash \{d\} \ \vee \\ & \exists e \in {}_2E \exists \sigma \in \Sigma \quad \mathcal{I}(e) = \{a\} \setminus\setminus \neg B \mid\!\!\vdash \{c\} \\ & \qquad \wedge\ \sigma a \in E \subseteq (\neg \sigma B)^c \ \wedge\ d = \sigma c]\}. \end{aligned}$$

[29] [REITER 1980], [REITER AND CRISCUOLO 1981]

(47.1) DEFINITION. *The Reiter-extensions of $S \subseteq \mathcal{D}$ are sets $E \subseteq \mathcal{D}_1$ such that $\Gamma_S(E) = E$, where for each $X \subseteq \mathcal{D}_1$, $\Gamma_S(X)$ is the smallest set satisfying*

(1) ${}_1S \subseteq \Gamma_S(X)$

(2) $\mathrm{Th}(\Gamma_S(X)) = \Gamma_S(X)$

(3) For each $d \in {}_2S$ and $\sigma \in \Sigma$, if $\mathcal{I}(d) = \{a\} \setminus\setminus B \mid\!\!\!\vdash \{c\}$, $\sigma a \in \Gamma_S(X)$, and $\neg\sigma B \cap X = \emptyset$, then $\sigma c \in \Gamma_S(X)$.

This definition rephrases in our language the definition of "extensions" given by REITER. The equivalence of this notion with our notion of admissible extension is seen as follows.

(47.2) THEOREM. *$S \lhd E$ iff ${}_2S = {}_2E$ and ${}_1E = \Gamma_S({}_1E)$.*

PROOF. (only if) Let $S \lhd E$. Clearly, if $e \in {}_2E$, then $e \in S$, since there are no reasons in $\mathcal{D}$ which support any elements of $\mathcal{D}_2$. Since $S \subseteq E$, this means ${}_2S = {}_2E$. Now note that E satisfies conditions (1)-(3) of the definition of Γ_S, so $\Gamma_S({}_1E) \subseteq E$, hence $\Gamma_S({}_1E) \subseteq {}_1E$. Next, since each element in $\mathcal{D}_1$ is trivially interpreted, ${}_2E \cup \Gamma_S({}_1E) \in QExts(S)$. But by the minimality of E among $QExts(S)$, $E \subseteq {}_2E \cup \Gamma_S({}_1E)$, hence ${}_1E = \Gamma_S({}_1E)$. (if) Suppose ${}_2S = {}_2E$ and ${}_1E = \Gamma_S({}_1E)$. Since each element of $\mathcal{D}$ is trivially interpreted, this means $E \in QExts(S)$, so $\Lambda_\omega(S, E) \subseteq E$. Now $\Lambda_\omega(S, E)$ satisfies (1)-(3) of the definition of Γ_S, so $E \subseteq \Lambda_\omega(S, E)$, hence $E = \Lambda_\omega(S, E)$ and $S \lhd E$. ∎

Note that in the logic of defaults, statements of the object language may not refer to defaults, while no such restriction is present in the theory of linguistic reasons.

§**48.** The next example is the logic of propositional deduction as cast by MCALLESTER in his TMS.[30] The principal motivation for this logic is the view that a reason $A \setminus\setminus B \mid\!\!\!\vdash C$ really means $A \wedge \neg B \supset C$, and that all sources for reasoned assumptions lie outside the ordinary logical system. This motivation is captured as follows as a special case of our formalization of REITER'S logic of defaults. As before, let $\mathcal{L}$ be a first-order logical language, and $\mathcal{D} = \mathcal{D}_1 \oplus \mathcal{D}_2$, where $\mathcal{D}_1 = \mathcal{D}_2$ is the set of closed wffs of $\mathcal{L}$. If $d \in \mathcal{L}$, we write ${}_1d$ for its occurrence in $\mathcal{D}_1$ and ${}_2d$ for its occurrence in $\mathcal{D}_2$, and as before, we write ${}_1S$ and ${}_2S$ to mean $S \cap \mathcal{D}_1$ and $S \cap \mathcal{D}_2$ respectively. The elements of $\mathcal{D}_1$ are called statements, the elements of $\mathcal{D}_2$ defaults. If ${}_1d \in \mathcal{D}_1$, $\mathcal{I}({}_1d) = \mathbf{P}\,\mathcal{D}$, and if ${}_2d \in \mathcal{D}_2$, $\mathcal{I}({}_2d) = \emptyset \setminus\setminus \{\neg {}_1d\} \mid\!\!\!\vdash \{{}_1d\} = \{S \subseteq \mathcal{D} \mid \neg {}_1d \notin S \supset {}_1d \in S\}$. In this special case of the logic of defaults, we have for each $d \in \mathcal{D}$ and $S \subseteq \mathcal{D}$,

$$\mathcal{J}(d, S) = \{E \mid d \in E \supset [d \in S \ \vee\ E - \{d\} \vdash \{d\} \ \vee\ d \in \mathcal{D}_1 \wedge [{}_2d \in {}_2E \ \wedge\ \neg d \notin E]\}.$$

This logic weakens the logic for default reasoning by removing all power of discussing assumptions and their motivations from the states themselves.

While this simple logical system may reflect MCALLESTER'S intent for the logic of propositional deduction, he actually proposes a specific deductively weaker system, a modal logic of sorts. In that system, $\mathcal{D}_1$ is the set of ground formulas of a first-order language extended with a modality **True** and $\mathcal{D}_2$ is the set of all formulas of the form **True**(p) or $\neg$**True**(p), where p is a ground formula of the non-modal first-order language. The idea is that reasons $A \setminus\setminus B \mid\!\!\!\vdash C$ are translated as **True**$(A) \wedge \neg$**True**$(B) \supset$ **True**(C), and this is carried out as follows. In contrast to the preceding, we now define admissible states to be sets of formulas closed under a complex set of inference rules, rules which together combine to produce something less than ordinary deducibility. Everything except deducibility is defined as before, and we define $\vdash$ to be the deducibility relation generated by the following abbreviations (explained immediately below).

1. $\ominus$**True**$(p \vee q)$ $\odot$ **True**(p) $\odot$ **True**(q)
2. $\ominus$**True**(p) $\odot$ **True**$(p \vee q)$
3. $\ominus$**True**(q) $\odot$ **True**$(p \vee q)$

[30] [MCALLESTER 1980]

4. $\mathbf{True}(p \wedge q) \ \odot \ \ominus\mathbf{True}(p) \ \odot \ \ominus\mathbf{True}(q)$
5. $\ominus\mathbf{True}(p \wedge q) \ \odot \ \mathbf{True}(p)$
6. $\ominus\mathbf{True}(p \wedge q) \ \odot \ \mathbf{True}(q)$
7. $\ominus\mathbf{True}(p \supset q) \ \odot \ \ominus\mathbf{True}(p) \ \odot \ \mathbf{True}(q)$
8. $\mathbf{True}(p \supset q) \ \odot \ \mathbf{True}(p)$
9. $\mathbf{True}(p \supset q) \ \odot \ \ominus\mathbf{True}(q)$
10. $\mathbf{True}(\neg p) \ \odot \ \mathbf{True}(p)$
11. $\ominus\mathbf{True}(\neg p) \ \odot \ \ominus\mathbf{True}(p)$

Here $\ominus$ and $\odot$ are "meta-negation' and "meta-disjunction." Each of the above statements abbreviates several inference rules corresponding to the logical structure of the meta-formula. For example, #7 above actually stands for the three rules

7.a $\mathbf{True}(p \supset q), \mathbf{True}(p) \vdash \mathbf{True}(q)$
7.b $\mathbf{True}(p), \neg\mathbf{True}(q) \vdash \neg\mathbf{True}(p \supset q)$
7.c $\mathbf{True}(p \supset q), \neg\mathbf{True}(q) \vdash \neg\mathbf{True}(p)$

Note the weakness of this $\vdash$ compared with ordinary deducibility, in that $\mathbf{True}(p \supset q), \mathbf{True}(\neg p \supset q) \nvdash \mathbf{True}(p)$. As with RMS, the exact conservation relation realized by MCALLESTER'S TMS is not clear, but seems related to the symmetric difference relation defined previously. Otherwise, the accuracy of this characterization of the weak logic of propositional deduction is easy to verify, as the above formulas merely paraphrase the definitions and explanations given by MCALLESTER, where we write $\mathbf{True}(p)$ and $\neg\mathbf{True}(p)$ for his $(p.\textit{true})$ and $(p.\textit{false})$.

§**49.** Our next example restates non-monotonic logic, one of the first formal treatments of reasoned assumptions. Non-monotonic logic is based on the idea of phrasing rules for making assumptions in terms of logical consistency of assumptions with other beliefs. This idea traces back to MCCARTHY and HAYES, who introduce but never develop modalities **Normally**, **Consistent**, and **Probably** for use in the rule $\mathbf{Normally}(p), \mathbf{Consistent}(p) \vdash \mathbf{Probably}(p)$.[31] Later, MCDERMOTT and the present author developed the idea by providing a formal theory involving statements of the form $p \wedge \mathbf{M}q \supset r$, where **M** here is a modality intuitively interpreted as logical consistency with other beliefs, that is, $\mathbf{M}p$ means $\neg p$ is not a consequence of current beliefs.[32] Since then there have been other developments by MCDERMOTT, STALLMAN, GABBAY, and MOORE. We discuss these below as well.

We cast the initial non-monotonic logic as follows. Let $\mathcal{D}$ be the set of sentences in a first-order language extended by the unary modality **M**. We let $\vdash$ stand for ordinary deducibility, and define

$$\mathcal{R} = \{S \subseteq \mathcal{D} \mid S = \mathrm{Th}(S) \ \wedge \ \forall d \in \mathcal{D}\,[\neg d \in S \ \vee \ \mathbf{M}d \in S]\}.$$

We make no nontrivial interpretations of elements of $\mathcal{D}$, so $\mathscr{S} = \mathcal{R}$. For each $d \in \mathcal{D}$ and $S \subseteq \mathcal{D}$ we define

$$\mathcal{J}(d, S) = \{E \mid d \in E \supset [d \in S \ \vee \ E - \{d\} \vdash \{d\} \ \vee \ \exists e \in \mathcal{D} \neg e \notin E \wedge d = \mathbf{M}e]\}.$$

Finally, we let $AExts = FGExts$, where $E \in FGExts(S)$ iff for each $e \in E$ there is a finite set $G \subseteq E$ and a well-ordering $<_G$ of G such that $e \in G$ and whenever $d \in G$, either $d \in S$ or there is a set $A <_G d$ with $A \vdash \{d\}$ or $d = \mathbf{M}f$ for some $f \in E^c$.

(49.1) THEOREM. *$S \lhd E$ iff $E = \mathrm{Th}(S \cup \{\mathbf{M}d \mid \neg d \notin E\})$.*

PROOF. (only if) Assume $S \lhd E$. By the admissibility of E, $\{\mathbf{M}d \mid \neg d \notin E\} \subseteq E$, since $S \lhd E$, $S \subseteq E$, and since $E = \mathrm{Th}(E)$, $\mathrm{Th}(S \cup \{\mathbf{M}d \mid \neg d \notin E\}) \subseteq E$. But if $e \in E$, then every grounding set G for e contains a proof of e from $S \cup \{\mathbf{M}d \mid \neg d \notin E\}$, so $E \subseteq \mathrm{Th}(S \cup \{\mathbf{M}d \mid \neg d \notin E\}$, hence $E = \mathrm{Th}(S \cup \{\mathbf{M}d \mid \neg d \notin E\})$. (if) Suppose $E = \mathrm{Th}(S \cup \{\mathbf{M}d \mid \neg d \notin E\})$. Clearly $E \in QExts(S)$. Furthermore, if $e \in E$, every proof of e from $S \cup \{\mathbf{M}d \mid \neg d \notin E\}$ is a grounding set for e in E, hence $E \in FGExts(S)$. ∎

[31] [MCCARTHY AND HAYES 1969]
[32] [MCDERMOTT AND DOYLE 1980]

Thus a simpler characterization of the admissible extensions of S is as the fixed points of deductively closing S together with all the "assumptions" of consistency not ruled out in the admissible extension. This characterization also shows the equivalence of our definition with that of MCDERMOTT and DOYLE, since this fixed-point formula is exactly their definition.

(49.2) COROLLARY. $AExts = \mu AExts$.

PROOF. Suppose $S \lhd E$, $S \lhd E'$, and $E \subseteq E'$. Let $A = \{\mathbf{M}d \mid \neg d \notin E\}$ and $A' = \{\mathbf{M}d \mid \neg d \notin E'\}$. Since $E \subseteq E'$, $A \supset A'$, so by the monotonicity of Th, $E' = \mathrm{Th}(S \cup A') \subseteq \mathrm{Th}(S \cup A) = E$. Hence $E = E'$. ∎

Unfortunately, this logic is too weak, in that it does not enforce the intuitive interpretation of **M** as consistency. For example, the set of axioms $\{\neg P, \mathbf{M}P\}$ is perfectly consistent in this logic, and its sole admissible extension contains no pair of contrary sentences. We can strengthen the theory to capture the intent by adding interpretations for each $d \in \mathcal{D}$ of $\mathcal{I}(\mathbf{M}d) = \{S \subseteq \mathcal{D} \mid \neg d \notin S \equiv S = \mathcal{D}\}$. With this added requirement, we remove from $\mathscr{S}$ all elements of $\mathcal{R}$ presenting an incoherent but not inconsistent notion of consistency for **M**, but lose the simple characterization of admissible extensions given in Theorem 49.1.

EXAMPLES. Here we use the strengthened logic.

(49.3) $S = \{\mathbf{M}P \supset P\}$, $\quad AExts(S) = \{\mathrm{Th}(\{\mathbf{M}P \supset P, P\} \cup \{\mathbf{M}P, \ldots\})\}$

(49.4) $S = \{\mathbf{M}P, \neg P\}$, $\quad AExts(S) = \emptyset$

(49.5) $S = \{\neg P, \mathbf{M}(P \wedge Q)\}$, $\quad AExts(S) = \emptyset$

(49.6) $S = \{\mathbf{M}P \supset Q, \neg Q\}$, $\quad AExts(S) = \emptyset$

(49.7) DEFINITION. *A set $S \subseteq \mathcal{D}$ has default form if every formula in S is either non-modal or has only non-iterated occurrences of* **M** *in the form $p \wedge \mathbf{M}q_1 \wedge \ldots \wedge \mathbf{M}q_n \supset r$, for some non-modal formulas $p, q_1, \ldots q_n, r$.*

(49.8) THEOREM. *If S has default form and $S \lhd E$, then for each $d \in \mathcal{D}$, $\mathbf{M}d \in E$ iff $\neg d \notin E$.*

PROOF. Suppose S has default form and $S \lhd E$. By the admissibility of E, $\mathbf{M}d \in E$ if $\neg d \notin E$, so suppose $\mathbf{M}d \in E$. Then $\mathbf{M}d$ must have a proof from $S \cup \{\mathbf{M}e \mid \neg e \notin E\}$. But since $\mathbf{M}d$ is not in default form, $\mathbf{M}d \notin S$, and similarly, $\mathbf{M}d$ cannot appear as a conclusion of Modus Ponens in such a grounding argument, hence $\mathbf{M}d$ must be an assumption, that is, $\neg d \notin E$. ∎

Unfortunately, a default form set S can be incoherent even if its natural translation S' in the linguistic reasons theory is coherent. REITER'S example is the set S containing just

$$\begin{array}{ccc} p_1 \wedge \mathbf{M}q_1 \supset q_1 & p_2 \wedge \mathbf{M}q_2 \supset q_2 & p_3 \wedge \mathbf{M}q_3 \supset q_3 \\ q_1 \supset p_2 & q_2 \supset p_3 & q_3 \supset p_1 \\ q_1 \supset \neg q_2 & q_2 \supset \neg q_3 & q_3 \supset \neg q_1. \end{array}$$

The translation of this, S', is a set of normal defaults, and so is coherent by Theorem 24.15 with extension $\mathrm{Th}(S')$. But S itself is incoherent. To see this, suppose $S \lhd E$. If $\mathbf{M}q_1, \mathbf{M}q_2, \mathbf{M}q_3 \notin E$, then $\neg q_1, \neg q_2, \neg q_3 \in E$, but these must be ungrounded, hence at least one assumption must be in E. E cannot contain two or more of these assumptions, since each rules out the previous one: for instance, if $\neg q_1 \notin E$, then $\mathbf{M}q_1 \in E$, hence $p_1 \supset q_1 \in E$, hence $\neg q_3 \supset q_1 \in E$, hence $\neg q_3 \in E$. But E cannot contain just one of these assumptions either, since one does not rule out both of the other two, thus allowing a second. Therefore S cannot be coherent.

While the unwanted admissible states can be removed from the original non-monotonic logic by the surgical addition of nontrivial interpretations for modal statements, the surgery leaves an ugly scar, in that we no longer have a simple characterization of the resulting states. Fortunately, MOORE has discovered a better solution, which, simply stated, defines admissible extensions of S as sets E such that

$$E = \mathrm{Th}(S \cup \{\mathbf{M}d \mid \neg d \notin E\} \cup \{\neg\mathbf{M}\neg d \mid d \in E\}).$$

(See [MOORE 1983].) This definition remedies the weaknesses of the original non-monotonic logic while retaining a simple characterization of admissible extensions. We reproduce this idea by redefining the general restriction $\mathcal{R}$ so that

$$\mathcal{R} = \{S \subseteq \mathcal{D} \mid S = \mathrm{Th}(S) \ \wedge \ \forall d \in \mathcal{D}[d \in S \equiv \neg\mathbf{M}\neg d \in S \ \wedge \ d \notin S \equiv \mathbf{M}\neg d \in S]\}$$

and by redefining $\mathcal{J}$ so that for each $d \in \mathcal{D}$ and $S \subseteq \mathcal{D}$,

$$\begin{aligned}\mathcal{J}(d,S) = \{E \mid d \in E \supset [d \in S \quad &\vee \quad E - \{d\} \vdash \{d\} \\ &\vee \quad \exists e \in \mathcal{D} \quad d = \mathbf{M}e \ \wedge \ \neg e \notin E \\ &\vee \quad \exists e \in E \quad d = \neg\mathbf{M}\neg e]\},\end{aligned}$$

with the corresponding redefinition of $FGExts$. If we read $\neg\mathbf{M}\neg$ as "in" and $\mathbf{M}\neg$ as "out," we see that MOORE'S requirement is to make admissible states omniscient about their contents, in that every admissible state S completely encodes the "in" and "out" sets, since $S = \{e \mid \neg\mathbf{M}\neg e \in S\}$ and $S^c = \{e \mid \mathbf{M}\neg e \in S\}$. MOORE suggests renaming non-monotonic logic as "autoepistemic logic" to recognize this sort of self-omniscience.

Other approaches were taken towards remedying the weaknesses of the initial non-monotonic logic, but they introduced other problems and so are less attractive than MOORE'S solution. MCDERMOTT tried strengthening the weak logic by adding axiom schema and inference rules for **M** similar to those seen in classical modal logic.[33] He considered systems corresponding to the modal logics **K**, **T**, **S4**, and **S5**, which employ the inference rules

MP : $p, p \supset q \vdash q$
Nec : $p \vdash \mathbf{L}p$

(here **L** abbreviates $\neg\mathbf{M}\neg$) and the axiom schema

Taut :	$\mathbf{L}(t)$	for all tautologies t
K :	$\mathbf{L}(p \supset q) \supset (\mathbf{L}p \supset \mathbf{L}q)$	
T :	$\mathbf{L}p \supset p$	
S4 :	$\mathbf{L}p \supset \mathbf{LL}p$	
S5 :	$\mathbf{M}p \supset \mathbf{LM}p$	

The resulting system including **Nec**, **Taut**, and **K** allows inference of $\mathbf{M}P$ from $\mathbf{M}(P \wedge Q)$, something not possible in the weak logic, and these additions make $\mathbf{M}P$ and $\neg P$ inconsistent. But the incoherence of $\{\mathbf{M}P, \neg P\}$ in the weak logic is seen as a lack of self-omniscience, so **T**, **S4**, and **S5** are added in one at a time to give the logic a description of its own sense of provability. No convenient characterization of the power of the systems of non-monotonic **T** and **S4** are known, and unfortunately in the extreme case, these additions trivialize the logic by making all reasons invertible in non-monotonic **S5**. That is, in the system with **Nec**, **Taut**, **K**, **T**, **S4**, and **S5** we can conduct the following proof:

1. $\mathbf{M}P \supset P$	hypothesis
2. $\mathbf{L}(\mathbf{M}P \supset P)$	**Nec**, 1
3. $\mathbf{LM}P \supset \mathbf{L}P$	**K**, 2, **MP**
4. $\mathbf{L}\neg P \supset \neg P$	**T**
5. $P \supset \mathbf{M}P$	4, tautologies
6. $P \supset \mathbf{LM}P$	5, **S5**, **MP**
7. $P \supset \mathbf{L}P$	6, 3, **MP**
8. $P \supset \neg\mathbf{M}\neg P$	7, rewriting
9. $\mathbf{M}\neg P \supset \neg P$	8, tautologies

Because of this, we cannot use the axiom $\mathbf{M}P \supset P$ to express a preference that P should be adopted before $\neg P$, since from this statement we can infer $\mathbf{M}\neg P \supset \neg P$, the opposite "preference." In fact, the motivations for the additional axioms seems to be to allow the logic to invert all of its reasoning by allowing the discussion of proof steps within the language, and as we saw earlier, theories of invertible reasons are bound to be trivial (an observation only made following MCDERMOTT'S work).

[33] [MCDERMOTT 1982A]

Yet another approach toward strengthening the weak logic was explored by GABBAY.[34] Rather than add in axioms specifically to remedy weaknesses, as in the modal logic extensions, GABBAY began with models for intuitionistic predicate calculus. In these models, the set of beliefs is viewed as a set of theorems monotone nondecreasing in number with the passing of time. At each temporal instant in such models, $\mathbf{M}P$ is interpreted to mean "it is consistent at this instant to assume that P is true." While these models permit a motivated development of natural axioms and inference rules, they unfortunately trivialize the logic, for $\mathbf{M}P \supset P$ is semantically equivalent with intuitionistic $\neg P \vee P$. Because of this equivalence, stating a "reason" like $\mathbf{M}P \supset P$ does not express any preference, but only requires that every model of the axioms begin with one of P or $\neg P$ being held true, without saying which one.

Realizing the inadequacies of the modal logic extensions of the weak non-monotonic logic, and discerning the need for non-invertible reasons that express preferences, STALLMAN proposed an extension of non-monotonic **S5** which employs a unary modality, **S**, interpreted as "should be a theorem" in contrast to **L**'s reading as "is a theorem."[35] All of the inference rules and axiom schema of non-monotonic **S5** are assumed, as is the new inference rule $\mathbf{S}p \vdash p$. This allows expression of defaults as statements of the form $\mathbf{M}p \supset \mathbf{S}p$.

It is unfortunate that STALLMAN'S thoughts have remained unpublished, for they are very insightful. The present author developed his ideas about ratiocinative desires as expressed in the attitudinal theories of reasoned assumptions above by the fortuitous simultaneity of his attempts to express the logic of reasoned deliberation in non-monotonic logic and his attempts to understand the varieties and possible improvements of non-monotonic logics, during which he realized the possibility of identifying the "should" modality of STALLMAN'S logic with the notions of ratiocinative intentions or desires. This identification led naturally to the idea that the formal constructions of non-monotonic logic might be plausibly motivated in decision-theoretic terms.

§**50.** To back away from this preoccupation with logically structured agents, we consider some elements of MINSKY'S K-line theory of memory.[36] Unfortunately, his theory involves many ideas beyond those we formalize, so our presentation is meager compared with his. One does not do justice to MINSKY'S conceptions to suppose the following representative of more than the simplest elements of his theory.

For MINSKY, the mind is composed of a set of "mental agents." Each mental agent can be either active or inactive, and states of mind are simply sets of active mental agents. We can identify the set of mental agents with the domain $\mathcal{D}$ of the agent, and consider sets $S \in \mathscr{S}$ to be the admissible sets of active mental agents.

The two specific sorts of mental agents we formalize here are K-lines and cross-exclusion networks. K-lines are mental agents that, when activated, cause the activation of some set of other mental agents. We formalize this by interpreting K-line agents as monotonic simple reasons. Specifically, for each K-line *KL* there is some set $A \subseteq \mathcal{D}$ such that $\mathcal{I}(KL) = \emptyset \setminus\setminus \emptyset \mid\!\!\sim A$.

A purely monotonic agent is not terribly interesting, and one source of non-monotonicity in this theory is that of cross-exclusion networks. These are sets of mental agents which are mutually inhibitory. Further, cross-exclusion networks facilitate "conflict resolution" by disabling or ignoring all members if two or more manage to become active despite their mutual inhibitions. This disabling allows activation of "higher-level" mental agents which can consider and resolve the conflict. We might formalize this by letting *CXN* be the mental agent representing a cross-exclusion network, $D = \{d_i\}_{i=1}^n$ the set of mutually inhibiting members, $E = \{e_i\}_{i=1}^n$ the indicators of which competitor wins out, and $\neg$*CXN* a mental agent representing the existence of an externally forced conflict. To get the desired behavior, we define $\mathcal{I}(CXN) = \emptyset \setminus\setminus \{\neg CXN\} \mid\!\!\sim D, \mathcal{I}(d_i) = \emptyset \setminus\setminus E - \{e_i\} \mid\!\!\sim \{e_i\}$ for each i, and assume the existence of a "watchdog" *WD* such that $\mathcal{I}(WD) = \{S \subseteq \mathcal{D} \mid [\exists i \neq j \leq n \quad e_i, e_j \in S] \supset \neg CXN \in S\}$. This interpretation of *WD* cannot be expressed as a single simple reason, although it can be expressed as a set of $n(n-1)$ simple reasons.

[34] [GABBAY 1982]
[35] [STALLMAN 1981]
[36] [MINSKY 1980]

Finally, states of mind persist until changes are forced by inputs. It is impossible to treat persistence in detail without first formalizing many parts of the theory too involved to discuss here, but we point out that persistence is closely related to conservation, and its formal treatment may well take the form of transition comparison relations for use by a conservative agent.

§**51.** Our final subject among theories related to reasoned assumptions is MCCARTHY'S notion of *circumscription*.[37] Circumscription has figured prominently in discussions of non-monotonic reasoning, for like theories of reasoned assumptions, it formalizes certain patterns of unsound inferences important in artificial intelligence. Unfortunately, previous discussions have never proved successful at satisfactorily relating the two notions, in spite of their common motivations and superficially similar formal treatments. This section attempts an explanation of this failure of understanding. We first present a description of circumscription in its own terms, as formulated by MCCARTHY. This is prelude to our main conclusion, that when closely examined, the notions of circumscription and reasoned assumption are almost entirely unrelated, both conceptually and formally. This need not be an unhappy conclusion, if one enjoys richness in one's subjects of study. We conclude the section by sketching some practical and theoretical aspects of agents in which the two notions are combined.

The idea of circumscription is that for each predicate occurring in a set of sentences in a logical language, one can construct an axiom schema which states that the only truths involving that predicate are those which follow from the original set of sentences alone. Put another way, the circumscription of some predicate states those conclusions which hold in all *minimal models* of the original set of sentences. For example, if the set of sentences states solely that red and blue are colors, circumscribing the sentences with respect to the predicate "color" produces a theory in which yellow is not a color. Since the sentences do not mention yellow, there are models of those sentences in which yellow is not a color. Some of those are models in which red and blue are the only colors, so one of the conclusions in the circumscription of the sentences is that yellow is not a color, indeed, that red and blue are the only colors. This sort of inference is non-monotonic because if the original sentences are augmented with the statement that yellow too is a color, the color-circumscription of the augmented set no longer contains the conclusion that yellow is not a color.

(51.1) DEFINITION. *Let A be a sentence (or conjunction of sentences) in a first-order language containing a predicate symbol $P(\mathbf{x}) = P(x_1, \ldots, x_n)$. Write $A(\Phi)$ for the result of replacing all occurrences of P in A by the predicate expression Φ. Then Circ(P, A), the circumscription of P in $A(P)$, is the sentence schema*

$$A(\Phi) \wedge \forall \mathbf{x}.(\Phi(\mathbf{x}) \supset P(\mathbf{x})) \supset \forall \mathbf{x}.(P(\mathbf{x}) \supset \Phi(\mathbf{x})).$$

We write $A \ \mu\vdash_P q$ iff $q \in \mathrm{Th}(\{$Circ$(P, A)\})$.

MCCARTHY observes that this definition can be extended to circumscriptions on two or more predicates simultaneously.

To illustrate this formalization, suppose we know only one red-haired person, our friend Jane. If we see someone looking like Jane in the crude sense of merely being red-haired, we might, via circumscription, assume that that person is Jane, she being the only person we know fitting that description. This inference is non-monotonic, of course, since if we now learn that Jane has an identical twin sister Joan, we can no longer conclude that anyone who looks like Jane is Jane. Expressed formally in terms of McCarthy's circumscription, this example might be translated as follows. We start with the set of axioms $A = \{$*red-haired*$($*Jane*$)\}$ and circumscribe on the predicate "*red-haired*." The circumscription of this predicate in A is the axiom schema

$$\Phi(\textit{Jane}) \wedge \forall x(\Phi(x) \supset \textit{red-haired}(x)) \supset \forall x(\textit{red-haired}(x) \supset \Phi(x)).$$

If we now substitute our only known instance of a red-haired person into this schema, that is, if we substitute the formula $x =$ *Jane* for $\Phi(x)$, we get

$$\textit{Jane} = \textit{Jane} \wedge \forall x(x = \textit{Jane} \supset \textit{red-haired}(x)) \supset \forall x(\textit{red-haired}(x) \supset x = \textit{Jane}).$$

[37] [MCCARTHY 1980], [DAVIS 1980]

The first two parts of this formula are true, and simplifying it leaves the resulting assumption or "default" $\forall x$(*red-haired*$(x) \supset x =$ *Jane*) which we can apply to any new person that looks like Jane (is red-haired). Yet this inference is non-monotonic, in that if we add the new axiom *red-haired*(*Joan*) to A, we can no longer draw any such identifying conclusion. At best, we can infer via another application of circumscription the less specific conclusion $\forall x$(*red-haired*$(x) \supset [x =$ *Jane* $\vee$ $x =$ *Joan*$]$).

(51.2) DEFINITION. *Let $M(A)$ and N(A) be models of the sentence A. We say that M is a submodel of N in P, writing $M \leq_P N$, if M and N have the same domain, all other predicate symbols in A besides P have the same extensions in M and N, and the extension of P in M is included in its extension in N. The model M is minimal in P iff $N \leq_P M$ only if $N = M$. The sentence A minimally entails q with respect to P, written $A \ \mu\models_P q$, iff q is true in all models of A that are minimal in P.*

(51.3) THEOREM (MCCARTHY-DAVIS). *If $A \ \mu\vdash_P q$, then $A \ \mu\models_P q$.*

PROOF. Let M be a model of A minimal in P. Let P' be a predicate satisfying the left side of $Circ(P, A)$ when substituted for Φ. By the second conjunct of the left side, P is an extension of P'. If the right side of the instantiated circumscription schema were not satisfied, P would be a proper extension of P'. In that case, we could get a proper submodel M' of M by letting M' agree with M on all predicates except P and agree with P' on P. This would contradict the assumed minimality of M. ∎

(51.4) THEOREM (DAVIS). *There are satisfiable sentences with no minimal models.*

PROOF. Let A be the conjuction of the following four sentences.

(1) $\forall x \exists y \ Succ(x, y)$
(2) $\exists y \forall x \ \neg Succ(x, y)$
(3) $\forall x, y, z \ [Succ(x, y) \wedge Succ(x, z) \supset y = z]$
(4) $\forall x, y, z \ [Succ(y, x) \wedge Succ(z, x) \supset y = z]$

Every model of A contains a submodel isomorphic to the natural numbers. But this submodel contains an infinite chain of sub-submodels corresponding to the sets of natural numbers exceeding k for each natural number k. Hence A has no minimal model. ∎

(51.5) THEOREM (DAVIS). *There are A and P such that $A \ \mu\models_P q$ but not $A \ \mu\vdash_P q$.*

PROOF. Let A be the conjunction of the following set of sentences.

(1) $\exists x \ Zero(x)$
(2) $\forall x, y \ [Zero(x) \wedge Zero(y) \supset x = y]$
(3) $\forall x \exists y \ Succ(x, y)$
(4) $\forall x, y \ [Succ(x, y) \supset \neg Zero(y)]$
(5) $\forall x, y \ [Zero(x) \supset Plus(x, y, x)]$
(6) $\forall x, y, z, u, v \ [Plus(x, y, z) \wedge Succ(y, u) \wedge Succ(z, v) \supset Plus(x, u, v)]$
(7) $\forall x, y \ [Zero(y) \supset Times(x, y, y)]$
(8) $\forall x, y, z, u, v \ [Times(x, y, z) \wedge Succ(y, u) \wedge Plus(z, x, v) \supset Times(x, u, v)]$
(9) $\forall x, y, z \ [(Zero(x) \vee Succ(y, x) \vee Plus(y, z, x) \vee Times(y, z, x)) \supset Number(x)]$

It is easily seen that there is a unique minimal model of A, namely the standard model of arithmetic. Hence $A \ \mu\models_{Number} q$ iff q is a true sentence of arithmetic. But the elements of $Circ(Number, A)$ are recursively enumerable, while the truths of arithmetic are not, hence there are circumscriptively true but underivable sentences. ∎

I have repeated these definitions and theorems virtually verbatim from their sources, both because they are worth knowing, and to emphasize the intimate connection they illustrate between the circumscriptive rule of inference and the notion of entailment in minimal models. This is important, for in my view circumscription is a natural and proper topic within the main tradition of mathematical logic. The heart of standard mathematical logic is the study of the entailment relation: When do the models of one set of sentences include the models of some other set of sentences? Analyzing these relations between sets of models leads naturally to analyzing the

structure of the class of all models, and as in most mathematical fields, there are natural orders relating the objects of study. In model theory, the model-submodel relation is one such order, so circumscription arises naturally when one studies entailment from the viewpoint of the model-inclusion order. Circumscription would arise naturally in mathematical logic even if no one cared about psychology, even if no one cared to mechanize intelligence. The notion of circumscription is logical, not psychological.

On the other hand, the notion of reasoned assumption is psychological, not logical, at least as far as standard mathematical logic is concerned. Consider the theory of simple reasons. This theory captures essentially all the principal motivations and characteristics of reasoned assumptions, but few logical notions are in evidence. Instead, the important notions are the psychological concepts of intention or desire, and of economic or decision-theoretic tradeoffs between simultaneously unsatisfiable ratiocinative desires and intentions. Questions of entailment and deducibility are not prominent; questions of utility and feasibility are. One can, of course, view state components as axioms and admissible extensions as sets of theorems, but the existence of interesting finite agents realizing these theories belies the identification, for no familiar logic has only a finite language, or has as little structure among its sentences as do some perfectly rigorous simple reasons agents.[38]

These disparate notions have been confused because of the circumstances in which they were developed. Theories of reasoned assumptions can take logical form, as seen in the linguistic reasons theory, and their first formal treatment was in non-monotonic logic, which as we observed previously, attempted to phrase reasoned assumptions in terms of logical consistency with sets of axioms. Since only logical terms appeared in that development, non-monotonic logic claimed logical status. Thus on one side, the properly psychological notions masqueraded as logical notions. But on the other side, the properly logical notion of circumscription found billing as a psychological notion, at least in artificial agents. Circumscription, of course, has application in certain psychologies, but that no more makes it a principally psychological notion than does the prevalence of carbon-based chemistry in human brains fit carbon for psychological prominence. It is instructive to compare the case of Modus Ponens. From a psychological point of view, circumscription and Modus Ponens are equally relevant, and equally foreign. Both are concepts from mathematical logic whose mechanizations find application in implementing certain psychological functions in certain agents. In this respect, circumscription is at a disadvantage because it is less mechanizable than Modus Ponens, and because much less is known about how and when to fruitfully apply it. But discretionary use of these techniques in implementing psychological functions does not make them crucial for psychological problems or make them psychological notions per se. There is really no more (and no less) need to "explain" the connection between circumscription and theories of reasoned assumptions than there is to "explain" the connection between Modus Ponens and theories of reasoned assumptions.

The issue is further complicated by the deductivists in psychology and the psychologists in logic. Deductivists are those who, in carciature at least, view all psychological problems as questions of formulating the appropriate logical axioms so that all mental activity can be phrased in terms of deductions from these axioms. As a claim about interesting psychologies, this view is either preposterous or trivial, depending on how one interprets it. As a methodology for how to conduct research in artificial intelligence, it is much less preposterous or trivial, but not without its problems. Psychologists in logic are those who, again in carciature, view logic as the study of the "laws of thought," a more general view than that taken in standard mathematical logic, and one which leads to all sorts of extensions of logic to incorporate psychological concepts. As a methodology, emphasizing logical investigation of psychological concepts, this view too has its merits, however distasteful the abuse of the term "logic" is to those with classical views. But as a thesis that there are laws including and beyond those of logic that must be common to the psychologies of all rational agents, the claim is unsubstantiated and very suspect, not only due to the diversity of imaginable psychologies, but also because the notion of a natural, universal standard of rationality is itself suspect.

It is important not to misconstrue this discussion of the psychological or logical centrality of certain notions as an attempt to mark boundaries between these fields, or to require labelling of all notions in terms of recognized disciplines. Topics in one field metamorphose into topics in the other quite frequently. But psychology's main aim is to study agents, their minds and their actions, and standard logic's main aim is to study truth itself. The two should not be identified lightly.

[38] Indeed, from the viewpoint of psychology, even the formal notion of axiom acquires nonlogical force, since the word word "axiom" derives from the Greek *axioma* and *axios*, words for worth or value and things thought worthy or valued.

Let us turn from distinctions to connections. On the practical side, almost nothing is known about when to circumscribe (in advance? when stuck?), what to circumscribe (all of the agent's beliefs? just a few?), how to circumscribe (mechanize mathematical induction?!), and when and how to retreat from circumscriptively obtained conclusions. The ideas extant (but mostly untried) run as follows. The preferred way of organizing logically-based agents these days is in terms of multiple, mutually referential theories; in the extreme, separate theories for each component for each concept, purpose, and activity of the agent.[39] In these contexts, circumscription seems suited to the formulation of defaults, which are then treated by the methods of reasoned assumptions. For example, one constructs a logical theory axiomatizing some concept, say rowboats. This "definitional" theory is usually incomplete, say by stating that the boat either can be rowed or lacks oars. One might employ circumscription to complete the theory, computing all the main conclusions that follow in which the boat lacks any problems.[40] These conclusions, such as the presence of oars, the soundness of the hull, etc., can then be added to the definitional theory in the form of default rules, making the assumptions whenever not specifically ruled out. Similarly, one might use circumscription "on demand" in problem-solving situations by routinely taking the statement of the problem and using circumscription to compute all the basic facts, for example definite lists of all objects and all predicate extensions, making those not already believed to be reasoned assumptions whose qualification is that nothing new is learnt about the problem formulation (new in the sense of not entailed by the initial formulation and hence in contradiction to the assumptions).

On the theoretical side, one can attempt to connect circumscription and reasoned assumptions in several ways. First, one can formulate the definition of something like inevitability in terms of circumscription. Here one has axioms stating the presence or absence of some state components, and other axioms giving the interpretations of state components and general restrictions (in other words, embedding the theory of reasoned assumptions in the agent's own language). Circumscribing these with respect to the "present" predicate on state components yields conclusions about what must hold in all minimal models of the axioms, about what components are inevitable given the initial components. Unfortunately, I do not know whether one can make this suggestion precise in any interesting way.

(51.6) QUESTION. *Assume a linguistic reasons agent is axiomatized in the suggested way, and suppose $S \subseteq \mathcal{L}$ has admissible extensions $\{E_1, \ldots, E_n\}$. Is it the case that circumscribing the axiomatization with respect to "present in the state" implies (or even is equivalent with) $\bigwedge E_1 \vee \ldots \vee \bigwedge E_n$?*

Another possible connection is to consider circumscribing a set of linguistic reasons with respect to all predicates at once. One expects this to contain more conclusions than the intersection of all admissible extensions, simply because the circumscription will try to complete all predicates, while the reasons in the state will only complete certain predicates with respect to certain instances.

(51.7) QUESTION. *Suppose $S \subseteq \mathcal{L}$ in a linguistic reasons agent is coherent. Does the circumscription of S with respect to all predicates properly include all sentences in $\bigcap AExts(S)$?*

More generally, we can ask for characterization of those psychologies in which important psychological notions match interesting logical notions "in the limit." The obvious candidate is minimal psychological entailment, even though it is not terribly important in the preceding development.

(51.8) QUESTION. *In what sorts of logically based psychologies is minimal psychological entailment the same as minimal entailment?*

Note that these questions are still problems for correct formulation rather than simply resolvable conjectures.

[39] See, for example, [WEYHRAUCH 1980], [KONOLIGE AND NILSSON 1980], [DOYLE 1980].

[40] There are sometimes computationally tractable ways of doing this, see for example [REITER 1982].

V. Conclusions

§**52.** We have come to the end of this story, even though there is still much to tell. In the preceding we explored the principal approach taken in artificial intelligence to the problem of acting with only incomplete information, that of employing rules of thumb for making and revising assumptions. We interpreted these rules of thumb as ratiocinative desires about when to be or not to be agnostic, how to resolve ambiguities, and when to abandon previous assumptions. We presented a mathematical framework in which each of the constituent ideas underlying applications of the approach could be individually introduced and analyzed. Within this framework we provided mathematical semantics for ratiocinative rules of thumb, which one might call "admissible state semantics" since the meaning of each reason is a set of sanctioned "admissible" states. The formal basis allows mathematical formulation and proof of many conceptions previously known only to folklore, and we stated and proved a number of these results. However, compared to the probabilistic approaches common in other fields, theories of reasoned assumptions are still near the beginnings of their development. I am acutely aware of the essential triviality of some of the results presented, but allow that a new field must begin somewhere. Nevertheless, some of the results are conceptually important, whatever their mathematical depth, for instance Theorem 27.16, the strong validity-optimality of admissible extensions. Hopefully the formal basis will permit deeper understanding of the issues involved, since now many questions of formulation can be treated technically rather than merely debated philosophically, and perhaps permit attack on the important practical questions about efficient mechanizations of conservative agents.

Many topics have been left for treatment elsewhere, either receiving no mention or only passing mention in the preceding. Among these come a development of the evolutionary theory in terms of the global structure of trajectory space; treatments of the abstract notions of reflection and conservatism used here concretely; relations of ideas discussed to CARNAP'S theory of probability, DACEY'S theory of conclusions, LEVI'S epistemic actions, SHAFER'S theory of evidence, and LEWIS'S theory of counterfactuals;[41] and development of practical revision systems satisfying attractive conservation specifications.

§**53.** While I have tried to draw connections between the methodologies of artificial intelligence and other fields, I have tried to avoid methodological debate within artificial intelligence itself, and hope a very few words of explanation will be tolerated. It is easy to misunderstand the work presented in the preceding, for its aims are somewhat different than those usual in artificial intelligence. The preceding does not attempt to solve a problem in the usual sense. It offers no new algorithms for realizing mental functions. Instead, its aim is to understand the problems and approaches already known. This involves the mathematician's methodological suspicion that the original formulation of a notion may not be the best one, and that better formulations may permit deep analysis where only complexity reigned before. I call this enterprise *rational psychology*, the investigation of psychology by reason alone, choosing this name after the example of rational mechanics. I say no more here, as details may be found elsewhere.[42]

[41] [CARNAP 1950], [DACEY 1978], [LEVI 1967], [SHAFER 1976], [LEWIS 1973]
[42] [DOYLE 1982], [DOYLE 1983A]

References

Birkhoff, G., 1967. *Lattice Theory*, third edition, Providence: American Mathematical Society.

Beltrametti, E. G., and Cassinelli, G., 1981. *The Logic of Quantum Mechanics*, Reading: Addison-Wesley.

Carnap, R., 1950. *Logical Foundations of Probability*, Chicago: University of Chicago Press.

Charniak, E., Riesbeck, C. K., and McDermott, D. V., 1980. *Artificial Intelligence Programming*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Dacey, R., 1978. A theory of conclusions, *Philosophy of Science* **45**, 563-574.

Davis, M., 1980. The mathematics of non-monotonic reasoning, *Artificial Intelligence* **13**, 73-80.

de Witt, B. S., and Graham, N., 1973. *The Many-Worlds Interpretation of Quantum Mechanics*, Princeton: Princeton University 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.

Doyle, J., 1983a. What is rational psychology?, Pittsburgh: Department of Computer Science, Carnegie-Mellon University.

Doyle, J., 1983b. Methodological simplicity in expert system construction: the case of judgments and reasoned assumptions, *AI Magazine*, to appear.

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

Frisch, A. M, and Allen, J. F., 1982. Knowledge retrieval as limited inference, *Proc. Sixth Conference on Automated Deduction* (D. W. Loveland, ed.), Lecture Notes in Computer Science # 138, Berlin: Springer-Verlag, 274-291.

Gabbay, D. M., 1982. Intuitionistic basis for non-monotonic logic, *Proc. Sixth Conference on Automated Deduction* (D. W. Loveland, ed.), Lecture Notes in Computer Science # 138, Berlin: Springer-Verlag, 260-273.

Gärdenfors, P., 1980. Epistemic importance and minimal changes of belief, Lund: Department of Philosophy, Lund University.

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

Goodman, N., 1973. The problem of counterfactual conditionals, *Fact, Fiction, and Forecast*, third edition, New York: Bobbs-Merrill, 3-27.

Goodwin, J. W., 1982. An improved algorithm for non-monotonic dependency-net update, Linköping: Software Systems Research Center, Linköping University.

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

Konolige, K., and Nilsson, N. J., 1980. Multiple-agent planning systems, *Proc. Conf. American Association for Artificial Intelligence*, 138-142.

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

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.

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

Lewis, D., 1980. A subjectivist's guide to objective chance, *Ifs* (W. L. Harper, R. Stalnaker, and G. Pearce, eds.), Dordrecht: Reidel, 267-297.

London, P. E., 1978. Dependency networks as a representation for modelling in general problem solvers, College Park: Department of Computer Science, University of Maryland, TR-698.

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

Martins, J. P., 1983. Reasoning in multiple belief spaces, Ph.D. thesis, Buffalo: Department of Computer Science, State University of New York.

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

McCarthy, J., 1980. Circumscription—a form of non-monotonic reasoning, *Artificial Intelligence* **13**, 27-39.

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.

McDermott, D., 1982a. Nonmonotonic logic II: nonmonotonic modal theories, *J. A. C. M.* **29**, 33-57.

McDermott, D., 1982b. Contexts and data-dependencies: a synthesis, Department of Computer Science, Yale University.

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

Minsky, M., 1962. Problems of formulation for artificial intelligence, *Proc. Symp. on Mathematical Problems in Biology*, Providence: American Mathematical Society, 35-46.

Minsky, M., 1963. 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. IFIP Congress*, 45-49.

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., 1980. K-lines: a theory of memory, *Cognitive Science* **4**, 117-133.

Minsky, M., and Papert, S., 1969. *Perceptrons: an introduction to computational geometry*, Cambridge: MIT Press.

Montague, R., 1963. Syntactical treatments of modality, with corollaries on reflection principles and finite axiomatizability, *Acta Philosophica Fennica* **16**, 153-167.

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

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

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.

Rabin, M. O., 1974. Theoretical impediments to artificial intelligence, *Information Processing 74*, Amsterdam: North-Holland, 615-619.

Rabin, M. O., 1976. Probabilistic algorithms, *Algorithms and Complexity: New directions and recent trends* (J. F. Traub, ed.), New York: Academic Press, 22-39.

Reiter, R., 1978. On reasoning by default, *Proc. Second Conf. on Theoretical Issues in Natural Language Processing*, 210-218.

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

Reiter, R., 1982. Circumscription implies predicate completion (sometimes), *AAAI-82*, 418-420.

Reiter, R., and Criscuolo G., 1981. On interacting defaults, *Proc. Seventh International Joint Conference on Artificial Intelligence*, 270-276.

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

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

Scott, D. S., 1982. A theory of domains and computability, lecture notes, Pittsburgh: Department of Computer Science, Carnegie-Mellon University.

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

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

Stallman, R. M., 1981. A new way of representing defaults, unpublished manuscript, Cambridge: Massachusetts Institute of Technology.

Stallman, R. M., and Sussman, G. J., 1977. Forward reasoning and dependency-directed backtracking in a system for computer-aided circuit analysis, *Artificial Intelligence* **9**, 135-196.

Szolovits, P., 1978. The lure of numbers: how to live with and without them in medical diagnosis, *Proc. Coll. Computer-Assisted Decision Making using Clinical and Paraclinical (Laboratory) Data* (B. E. Statland and S. Bauer, eds.), Tarrytown: Technicon, 65-76.

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

Thompson, A., 1979. Network truth maintenance for deduction and modelling, *Proc. Fifth International Joint Conference on Artificial Intelligence*, 877-879.

Tukey, J. W., 1960. Conclusions vs. decisions, *Technometrics* **2**, 423-433.

Turner, R., 1981. Counterfactuals without possible worlds, *J. Philosophical Logic* **10**, 453-493.

Van Frassen, B. C., 1980. A temporal framework for conditionals and chance, *Ifs* (W. L. Harper, R. Stalnaker, and G. Pearce, eds.), Dordrecht: Reidel, 323-340.

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

Zadeh, L., 1975. Fuzzy logic and approximate reasoning, *Synthese* **30**, 407-428.

Table of Symbols

Symbols appearing in the text are listed below in three groups. First are ones in the roman alphabet (alphabetically), second are ones in the greek alphabet (alphabetically), and third are other symbols. Some symbols appear more than once to avoid confusions about ambiguities of classification. The text also standardly uses the letters a, b, c, d, e, f, g to mean elements of $\mathcal{D}$, the letters A, B, C, D, E, F, G, S to mean subsets of $\mathcal{D}$, and $\hat{E}, \hat{S}$ to mean sets of subsets of $\mathcal{D}$.

$\mathcal{AE}$	*A posteriori* extent
$AExts$	Set of admissible extensions
Alt	Set of alternative states in backtracking
B^c	Complement of B in $\mathcal{D}$, i.e. $\mathcal{D} - B$
$\mathcal{CE}$	Conditional extent
$\mathcal{D}$	Domain of an agent's states
$\mathcal{E}$	Relative frequency of element in distribution
$Exts$	Set of extensions of a set
E	Extended transition function (Epsilon)
$FGExts$	Set of finitely grounded extensions
$GExts$	Set of grounded extensions
$\mathcal{I}$	State admissibility interpretation function
$\mathcal{J}$	Extension admissibility interpretation function
L	"Provable" modality
$\mathcal{L}$	A logical language
M	"Consistent" modality
m	Measure function
N	Natural numbers
NP	Class of non-deterministic polynomial time computable functions
$\mathcal{O}$	Complexity order class
P	Power set
P	Class of deterministic polynomial time computable functions
#P	Class of non-deterministic polynomial time countable functions
$\Pr$	Probability
$\Prh$	Probability of coming to hold a component
$\mathcal{Q}$	Set of component-admissible states
$QExts$	Set of component-admissible extensions
R	Real numbers
$\mathcal{R}$	Global restriction on admissible states
S	STALLMAN'S "Should" modality
$\mathscr{S}$	Set of admissible states
$\mathcal{T}$	Trajectory space of a system
Th	Closure of a set with respect to $\vdash$
$\mathcal{U}$	Universe of a set
$\mathcal{V}$	Set of valid reasons
∂	Kernel transition function
Δ	Full transition function
$\triangle$	Symmetric set difference
E	Extended transition function
Λ	Elements generated by all ranks
Λ_α	Elements generated by rank α

μ	Minimization operator
$\mu\vdash$	Circumscriptive deducibility relation
$\mu\models$	Minimal entailment relation
$\mu\mid\!\models_S$	Minimal psychological entailment
$\mu\mid\!\sim_S$	Minimal psychological derivability
ν	Transition minimization operator (“nearest”)
Σ	Set of closed term substitutions
$<_G$	Well-ordering of G
$\preceq$	Comparison of transition sizes
$\preceq_S$	Comparative similarity relation about S
$\triangle$	Symmetric set difference
$\lhd$	Extensions of sets
$\trianglelefteq$	Admissible extensions
$\ulcorner\ \urcorner$	Quasi-quotes
$A \setminus\setminus B \mid\!\vdash C$	Reason interpretation
$\vdash$	Deducibility relation
$\mu\vdash$	Circumscriptive deducibility relation
$\models$	Entailment relation
$\mu\models$	Minimal entailment relation
$\mid\!\sim$	Psychological arguability relation
$\mid\!\models_S$	Psychological entailment relation in S
$\mid\!\sim_S$	Psychological derivability relation
$\mu\mid\!\models_S$	Minimal psychological entailment
$\mu\mid\!\sim_S$	Minimal psychological derivability
$\oplus$	Direct (disjoint) sum
$\times$	Direct (Cartesian) product
$\ominus$	“Meta-negation,” see §48
$\odot$	“Meta-disjunction,” see §48
#P	Class of non-deterministic polynomial time countable functions