Reset reproduction of an article appearing in *Fundamenta Informaticae* Vol. 20, No. 1–3 (Spring 1994), pp. 35-73. Resetting corrects one error and changes fonts and pagination. Resetting prepared from submission of November, 1993. Submission and correction ©1993, 2006 Jon Doyle. All rights reserved. Author's current address: NCSU Department of Computer Science, Raleigh, North Carolina, USA.

Reasoned Assumptions and Rational Psychology

Jon Doyle DOYLE@LCS.MIT.EDU
Massachusetts Institute of Technology, Laboratory for Computer Science
Cambridge, Massachusetts 02139, U.S.A.

Abstract

Logical epistemology unduly sways theories of thinking that formulate problems of nonmonotonic reasoning as issues of nondeductive operations on logically phrased beliefs, because the fundamental concepts underlying such reasoning have little to do with logic or belief. These formulations make the resulting theories inappropriately special and hide the characteristic structures of nonmonotonic reasoning amid many unrelated structures. We present a more direct mathematical development of nonmonotonic reasoning free of extraneous logical and epistemological assumptions, and argue that the insights gained in this way exemplify the benefits obtained by approaching psychology as a subject for mathematical investigation through the discipline of *rational psychology*.

For Joseph A. Schatz, teacher and friend

1. Reasoning, logic, and psychology

Nonmonotonic reasoning, the study of making and revising assumptions in a reasoned or principled way, needs little introduction in artificial intelligence today thanks to years of extensive exposition, analysis, and application. In spite of an admirable history of progress, however, the subject stands in need of some rethinking and redirection as the strengths and limitations of the accepted theories become clearer. This paper seeks to further this rethinking and redirection by presenting the foundations of nonmonotonic reasoning through a mathematical and philosophical approach closer to the concepts and methods of modern physics and rational mechanics than to the standard formulations of artificial intelligence. I believe these concepts and methods, which seek to find the most appropriate means for describing and understanding psychological structure and behavior, will prove productive for rethinking other parts of artificial intelligence as well. This introduction thus attempts to set out some of the motivations for this rethinking and to motivate the methods underlying the formal treatment.

This paper celebrates the twentieth volume of *Fundamenta Informaticae*. The year of its writing (1993) also marks the twentieth anniversary of my involvement in the field of artificial intelligence; the fifteenth anniversary of the appearance of the original nonmonotonic logic (McDermott & Doyle, 1980); and the tenth and fifth anniversaries (respectively) of the appearance of my mathematical monograph (Doyle, 1983c) and my foundational monograph (Doyle, 1988), from which the present paper derives and upon which it improves, and looking back from these anniversaries has led me to include some personal interpretations of their history in this introduction.

1.1 Nonmonotonic reasoning

Though the reader will likely find the notion familiar, a few words about the term "nonmonotonic reasoning" should aid in understanding the discussion to follow.

Intuitively speaking, nonmonotonic reasoning refers to reasoning involving nonadditive changes in beliefs, preferences, intentions, and other mental attitudes. The intuitive notion, however, is meaningless on its own because reasoning is an activity, and activities are not inherently monotonic or nonmonotonic; any monotonicity and nonmonotonicity of reasoning must be relative to how we view the reasoning in terms of aspects of mental states. In the usual usage in theoretical artificial intelligence, one views mental states as consisting of sets of mental attitudes and reasoning as a process that fills out and changes these sets over time. One can thus identify two very different senses of nonmonotonicity of reasoning: *temporal* nonmonotonicity, in which mental attitudes may appear and vanish over time, and *logical* nonmonotonicity, in which filling out larger sets of attitudes may yield fewer conclusions than filling out smaller sets. Mathematically, temporal nonmonotonicity compares mental attitudes as time increases, while logical nonmonotonicity compares consequences as mental attitudes increase.

Temporal nonmonotonicity may occur routinely and unexceptionally, for example through direct temporal variation of mental attitudes by perceptual or cognitive systems that add and subtract attitudes to reflect changes or anticipated changes in the world (as might happen if some of the attitudes describe the contents of the retina). Logical nonmonotonicity may occur because the reasoner derives some attitudes as conclusions from others as long as the right circumstances obtain. In the canonical example, the reasoner infers that Tweety flies from the information that Tweety is a bird, but not from the information that Tweety is also a penguin, information that defeats or undercuts the usual conclusion. In general, however, the division between these two forms of nonmonotonicity is not sharp, as one may draw conclusions over time to convert logical nonmonotonicity into temporal nonmonotonicity, or replace mental simulations with atemporal reasoning or logics to convert temporal nonmonotonicity into logical nonmonotonicity (cf. (Makinson & Gärdenfors, 1991)).

Theorists and practitioners in artificial intelligence recognized the need for logically nonmonotonic reasoning early on, motivated by problems of reasoning about knowledge and actions, by the desire to make plausible commonsense inferences, and by the desire to speed problem-solving searches by making quick decisions about where to search that would yield information useful for guiding the search even if proven wrong. They suggesting ways of expressing nonmonotonic reasoning rules (e.g., (McCarthy & Hayes, 1969; Sandewall, 1972)) and implementing reasoning systems that performed versions of these (e.g., (Sussman, Winograd, & Charniak, 1971)), but rigorous and formal theories appeared later, for unlike ordinary logic, in which one takes contradictions to indicate flawed axioms, useful commonplace rules of nonmonotonic reasoning can provide conflicting conclusions in some cases, conflicts that call for adjudication, perhaps case by case, rather than for abandonment or revision of the conflicting rules. In another canonical example, the reasoner infers that Nixon is a pacifist because Nixon is a Quaker, but also infers Nixon is not a pacifist because Nixon is a (US) Republican, and has to decide which of these reasonable conclusions to accept while keeping the rules that led to them. The early proposals offered no precise ways of treating such conflicts, even when anticipated, as identifying coherent notions of nonmonotonic conclusions proved a perplexing task.

I formulated perhaps the first rigorous solution to this problem in 1976 as the two fundamental principles of my original reason maintenance system or RMS (Doyle, 1976, 1979) (renamed so from “truth maintenance system” or TMS in (Doyle, 1980)), which introduced the now-familiar notion of nonmonotonic justification. (Some may also consider McCarthy’s (1977) probably contemporaneous early notion of circumscription a solution to this problem, or even credit the older logical theory of implicit definition (Doyle, 1985).) The RMS represents mental attitudes (or other representational or procedural items) by structures called *nodes* that the RMS labels as either *in* or *out* (of the current state). The RMS also records sets of *justifications* or *reasons* for each node, most of which express simple boolean combinations of the labelings of nodes we denote as "$A \setminus\!\!\setminus B \parallel\!- c$" and read as "$A$ without B gives c", meaning that the node c should be *in* if each node in the set A is *in* and each node in the set B is *out*. The RMS then seeks to construct labelings for the nodes from these justifications, labelings that satisfy two principles: a "stability" principle of labeling each node *in* if and only if one of its reasons is *valid* in the labeling (i.e., expresses hypotheses "A without B" that match the labeling), and a "groundedness" principle demanding that labelings provide each node labeled *in* with a noncircular argument in terms of valid reasons. The structure for justifications given above makes both of these principles perfectly unambiguous. Indeed, these principles convert nonmonotonic reasoning tasks into problems for algorithmic analysis, and different versions of RMS explored different graph-theoretic techniques for analyzing systems of nodes and justifications.

1.2 Logical formalizations

The fundamental RMS principles led, in time, to a variety of formalizations of the stability and groundedness notions. The initial and most abidingly popular formalizations clothed these principles in logical garb: nonmonotonic logic (McDermott & Doyle, 1980) and the logic of defaults (Reiter, 1980), together with the circumscription rule of inference (McCarthy, 1977, 1980), which as a class of inference operators rather than a "logic" has a somewhat different character from nonmonotonic and default logics. Each of these theories formalizes nonmonotonic reasoning by encoding groundedness and the presence and absence of knowledge in terms of logical provability and unprovability, or in terms of logical consistency instead of provability. For example, the simplest transcription of the canonical example into nonmonotonic logic translates the inference as the implication $b \wedge \neg L \neg f \ \rightarrow\ f$, where L stands for the provability modality, b stands for Tweety's birdness, and f for Tweety's flying. A similar transcription of RMS justifications translates them into the form $a_1 \wedge \ldots \wedge a_m \wedge \neg L \neg b_1 \wedge \ldots \wedge \neg L \neg b_n \rightarrow c$.

Formalizations, as mathematical characterizations of ideas, may be either good or bad characterizations. The initial formalizations of nonmonotonic reasoning, along with their monotonously logical subsequent variants, improvements, and extensions, proved very good as ways of making the theoretical problems of logically nonmonotonic reasoning both interesting and accessible to a wider audience than artificial intelligence theoreticians. This accessibility and advertisement encouraged the involvement of many brilliant thinkers in the problems of artificial intelligence, and produced a large and vigorous literature that has significantly increased our understanding of nonmonotonic reasoning. It seems doubtful we would understand as much today had these formalizations not been developed and explored as they have.

At the same time, the logical formalizations proved very bad as conceptual characterizations of nonmonotonic reasoning because the fundamental concepts of nonmonotonic reasoning have little to do with the concepts of logic, which in these formalizations obscure and mislead one from attending to the concepts of interest. Understand that this does *not* mean the logical formalizations have no value, even as conceptual characterizations, for a bad formalization may serve well enough. But the logical formalisms deserve reconsideration because they highlight and enshrine at their core things essentially unrelated to nonmonotonic reasoning. We enumerate only three of the most important inappropriate aspects of the logical formalizations.

In the first place, the logical formalizations convert what in many systems is a fast and computationally trivial check for presence and absence of attitudes into a computationally difficult or impossible check for provability, unprovability, consistency or inconsistency. This inaptness seems especially galling in light of the initial problem-solving motivations for nonmonotonic assumptions, for which assumptions served to speed inference, not to slow it (cf. (de Kleer, Doyle, Steele, & Sussman, 1977; Ginsberg, 1991)).

In the second place, the logical formalizations impede development of realistic approaches to reasoning about inconsistent information. Standard logics make inconsistent theories trivial, and the use of logical provability or consistency to encode inferential permissions and guidance means that nonmonotonic theories must be consistent as well in order to be useful. But in practice, reasoners must deal with inconsistent information all the time. They may try to remove some inconsistencies when they deem the inconsistencies important enough to warrant the effort, but even so may take time to remove them, occasionally even a long time, and must keep operating reasonably throughout this process. A theory of reasoning basing the very definition of mental state on logical consistency requirements makes representation of such processes impossible.

In the third place, the logical formalizations encourage the view that nonmonotonic rules and defaults express beliefs or factual information, even though reasoning, and nonmonotonic reasoning in particular, may involve desires, intentions, and other mental attitudes besides belief (cf. (de Kleer et al., 1977; Doyle, 1980; McCarthy & Hayes, 1969; McDermott, 1978)). Now no one disputes that some nonmonotonic rules carry or presume some beliefs, but taking this special case as the general one has proven very misleading. For example, some theories attempting to provide guidance about choosing from among the possible interpretations of complex sets of nonmonotonic rules interpret these rules as qualitative statements about highly likely conditions. While high conditional probabilities of the conclusions may motivate adoption of some defaults and thus deserve attention as a special case, this interpretation does not even make sense for rules reasoning with other mental attitudes, and might indicate the wrong conclusions for all we know now. But such theories have been proposed as theories of nonmonotonic reasoning in general, not as theories of special cases. Nonmonotonic theories based on logical consistency provide another example with the same character. These theories convert reasoning about reasoning generally into reasoning about the consistency of beliefs. Now while one may motivate a number of reasonable principles for guiding reasoning about beliefs in terms of avoiding inconsistency, these principles simply do not apply to reasoning about desires and some other attitudes. The conclusions thus motivated hardly seem appropriate for general use without much further argument.

It is tempting to point the finger of blame for these mistakes elsewhere, even back to McCarthy and Hayes' (1969) original brief suggestion, but I bear much responsibility for these misleading formalizations due the initial nonmonotonic logic I developed with Drew McDermott in 1978 (McDermott & Doyle, 1980). I well understood the distinction between the underlying RMS ideas and the logical encoding, and was quite pleased when Reiter's (1980) default logic later remedied a number of the unsatisfying characteristics of our original nonmonotonic logic through an approach closer to the RMS. Nevertheless, I was for years as encouraged as anyone in thinking that some logic would eventually provide the right formulation for these ideas—at least until I undertook in 1981 to develop logics capturing the RMS approach even more precisely. I soon found a great variety of possible schemes of reasoning, which suggested that the fundamental ideas were best isolated and pursued independently rather than attempting to formulate each variant as a special quirky logic. Mathematical classification seemed more appropriate than philosophical or metaphysical proposal, especially as many of the dozens (if not hundreds) of the schemes imaginable seemed well suited to some special purpose, rather than some logic dominating the rest.

1.3 Rational psychology

Eventually it became clear that in spite of the great progress made through exploiting logical tools and theories in artificial intelligence, one cannot expect *a priori* that the most appropriate theory of psychology should have much to do with logic or should necessarily make extensive use of logical concepts. Psychology, as a subject for investigation, includes the study of many aspects of thought, feeling, and behavior, while standard logic mainly idealizes only part of one sort of mental activity, reasoning about facts or beliefs, and clearly does not address notions like intent, desire, and preference, nor even the relations between such attitudes and reasoning in terms of beliefs (to say nothing about its silence on love, fear, and other feelings). This is not to say one cannot cast parts of these larger theories in logical terms too; this enterprise goes back to Aristotle, and has been investigated diligently in modern philosophical logic to good effect. But even here, one finds no reason *a priori* to suppose that ordinary logic provides the most appropriate basis for these investigations. The beautiful theories of modern mathematical logic constitute a triumph of conceptual analysis; but while logic is great, it isn't psychology, and to use G. A. Miller's (1986) vivid metaphor, instead of carving up the subject of psychology at its joints and clearly revealing the structure of psychology's conceptual components, standard uses of logic may in fact dismember psychology, carving it up into ill-shapen lumps that reveal little of the attraction of the subject.

The rational mechanics of Newton and its modern renewal by Truesdell (1958, 1977) and his contemporaries provide a model for a more productive approach. Rational mechanics is a part of mathematics, the conceptual investigation of mechanics. "Rational" here indicates investigations based on reason alone, rather than on experiment, engineering, or computation, the rational analysis of the concepts and theories whose applicability and feasibility are studied in experimental, engineering, and computational projects. We call the corresponding part of mathematics devoted to the study of psychology *rational psychology* (Doyle, 1983d) (a term used by James (1892) and earlier writers to refer to philosophical psychology). Rational psychology is not the study of rational agents, but instead the mathematical approach to the problems of agents and their actions, whether one thinks these agents and actions rational or irrational. It aims to understand psychological ideas through mathematical classification of all possible minds or psychological systems, to describe and study mental organizations and phenomena by the most fit mathematical concepts, seeking the best (most appropriate, illuminating, edifying, powerful, . . .) ways of formalizing psychology. It excludes the problem of identifying important psychological phenomena except as a byproduct of organizing ideas about psychologies into a coherent mathematical whole.

The mathematical formulation of nonmonotonic reasoning presented below resulted from this shift in focus from logic to conceptual analysis, with many of the central concepts and results of the new formulation first circulated in my monograph (Doyle, 1983c), a dense exposition short on explanation that revealed the mathematical structure of the subject but doubtless exceeded the tolerance of most potential readers. That monograph introduced the notion of *simple reason*, which resembles a RMS justification or a propositional default in the logic of defaults, but which appears as a conclusion as well as as a rule, and which so also resembles one of Minsky's "K-lines" (Minsky, 1980). Simple reasons formed the focus of the original exposition, which deliberately presented the core of the development *twice*—once for simple reasons alone (no mention of any logical structure), and then again for states closed with respect to a compact abstract deducibility relation—to drive home the irrelevance of even minimal logical notions to the central mathematical structures of reasoned assumptions. That presentation also deviated from logic in embracing a wide conception of semantics and meaning as the theory of pure designation, without requiring the compositionality of meanings usually demanded by logical theories. It also hewed close to the RMS conception in making no assumptions that reasoned entities represented beliefs, or that presence or absence amounted to consistency or inconsistency, or that contrary beliefs should not be held indefinitely, or that any mental entities were inherently contradictory with any others. At the same time, it followed logic in aiming to describe reasoning agents regardless of computability, in seeking to describe the mind of God as well as the mind of Man. The present exposition also draws on my later monograph (Doyle, 1988), which provided better development of the motivation, better forms for some of the concepts, and better notation.

This paper omits treatment of many important topics addressed in (Doyle, 1983c) and (Doyle, 1988), in their numerous sequelae, and to some extent in (Doyle, 1980). We have omitted most discussion of the motivations for nonmonotonic reasoning in particular and the basic structures of mental architectures in general; of the nature of meaning and the structure of semantical theories; of "psycho-logics" derived from the state spaces of agents; of uniformly defeasible reasons; of the theory of denials and contradictions; of logical encodings of reasoning; of psychological attitudes; of conservatism and other topics concerning the evolution of mental states; of probabilistic constructions over trajectories and their relation to reasoning and the strength of mental attitudes; of the failings of practical systems like the RMS; of social, economic, and political structures within minds; of most connections with ideas from economics; and of reflection, deliberation, and action. The mathematical formulations of these topics presented in the earlier works sometimes need emendation, but generally fit well into the framework elaborated here.

This paper seeks to present the fundamental concepts and results about nonmonotonic reasoning in a setting free of unnecessary logical ornament. Section 2 introduces the notion of framings as different ways of viewing mental states, while Section 3 presents the underlying tools for describing the constitution or special structure of mental states, including the notion of satisfaction system used to express half of the RMS stability principle. Section 4 introduces simple reasons in semantic terms, as elements of mental states bearing certain constitutive meanings rather than as exhibiting specific structures, and develops canonical descriptions for them. Section 5 introduces constitutions for reasoning that express the other half of the RMS stability principle as well as the RMS groundedness principle, while Section 6 analyzes some of the structure of reasoned states.

2. Framings

We focus in this paper on aspects of mental states that shape or guide the course of reasoning. We do not presume to possess a full characterization of instantaneous mental states, nor do we make special assumptions about the nature of mental states. We instead cast the discussion instead in terms of representations of instantaneous states (mental or otherwise) that distinguish the structures of interest and focus on properties of structures with such representations. In doing so, we take care to differentiate between the activities of the reasoner and our descriptions of such actions, and to identify structures of reasoning in terms of their role in the reasoner's activities rather than in terms of how we represent them. That is, how we choose to talk about the reasoner and its reasoning should not affect what these things actually are.

We write $\mathcal{S}$ to denote the set of all possible instantaneous states of the agent or agents in question: in particular, the set of states reflecting all and only those distinctions relevant to the reasoning activities of interest. Ordinarily we know more ways of distinguishing states, and conceive of a larger set Ω of all possible instantaneous states, so that each state in $\mathcal{S}$ covers one or more states in Ω. In practice, we have no direct access to or complete knowledge of $\mathcal{S}$ (or Ω), and must instead work with various representations of $\mathcal{S}$ chosen to reflect the essential characteristics of $\mathcal{S}$ in a way amenable to reasoning and analysis. We call such faithful representations *framings*.

Definition 1 (Framings) *A* framing *of $\mathcal{S}$ is a one-to-one function $\phi : \mathcal{S} \rightarrow \mathcal{S}'$ taking states in $\mathcal{S}$ to representations in a set $\mathcal{S}'$. We call ϕ* exact *just in case it is onto as well.*

A framing thus provides a unique representation of each state, directly representing the chosen space of states without blurring distinctions or introducing new ones. It ensures that $\mathcal{S}$ captures only distinctions of interest since states in Ω differing only in irrelevant ways map to the same representation, and hence must be identified in $\mathcal{S}$. An exact framing, moreover, indicates an isomorphism $\mathcal{S} \cong \mathcal{S}'$, so that every difference among representations indicates a difference among states.

Two common types of framings provide the basic tools of our discussion: *multiattribute* and *elemental* framings. Multiattribute framings describe complex worlds in terms of several attributes at once, mapping each state s to a vector $\phi(s) = (\phi_1(s), \phi_2(s), \ldots)$ of attribute values, where the component attribution functions $\phi_i : \mathcal{S} \rightarrow A_i$ map $\mathcal{S}$ to component attribute spaces A_i. For example, common framings of physiological states assign vectors of height, weight, blood pressure, heart rate, etc. to each person. Elemental framings, on the other hand, describe states as consisting of sets of *state components*. An elemental framing of $\mathcal{S}$ over a set of components $\mathcal{D}$ is then a function $\phi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{D})$, representing each state s by a set $\phi(s) \subseteq \mathcal{D}$. For example, we often frame mental states in terms of a set of beliefs, desires, intentions, and other attitudes. Elemental framings may of course be viewed as particular multiattribute framings, with each state component corresponding to a distinct boolean attribute taking on the two values "present" and "absent". Conversely, we may view any strictly boolean multiattribute framing as an elemental framing. But conceptually and notationally the set-theoretic view seems more apt in many cases.

Elemental framings are especially natural in artificial intelligence, which often constructs agents with mental states consisting of sets of representations (symbolic or otherwise), for example, sets of statements in a logical language, sets of data-structures, or sets of active "mental agents" in a society of mind (Minsky, 1980). In such cases, we may choose trivial elemental framings that simply consist of the representations explicitly contained in the state; that is, in such cases, elemental framings need not simply provide ways of viewing states, but may actually provide the states themselves.

3. Constitutions

As a practical matter, one ordinarily chooses attributes or state components on the basis of considerations independent of the details of the full state space, indeed, independent of the specific agent in question; one usually chooses these descriptors to facilitate characterization of agents in some general class. Consequently, convenient and useful multiattribute and elemental framings rarely prove exact, that is, not every set of state components represents a possible state of the agent. This inexactness poses problems, both for external analysts and for deliberative agents reasoning about their own states and actions, since reasoning that considers every attribute vector or set of state components a possible state will err in some cases. To avoid these errors, we need to distinguish those representations that actually correspond to states from those that do not.

One way to distinguish representations of actual states from imposters is to describe a *constitution* over the set of possible representations, a set of "laws" or restrictions on the constitutionally *admissible* combinations of attribute values or state components. For example, a constitution for an ideal psychological agent framed in terms of standard mental attitudes might restrict framings to those in which the set of beliefs is always deductively closed. We add these restrictions until only representations of actual states remain, until the framing between states and the constitutionally admissible representations is exact. While one may construct constitutions for mental states as a way of specifying the design of some agent, we here do not seek to prescribe some psychology, only to identify, understand, and classify special psychologies, with particular attention to the ones we find most interesting. These special mental constitutions thus correspond directly to the constitutive assumptions used in rational mechanics to identify special classes of physical materials (cf. (Truesdell, 1977)).

3.1 Bounding systems

At their most abstract, constitutions simply set out the image $\phi(\mathcal{S})$ of the framing under consideration. The map ϕ indicates an isomorphism of $\mathcal{S}$ with $\phi(\mathcal{S})$, so the constitution $\phi(\mathcal{S})$ permits one to ignore $\mathcal{S}$ itself and simply work with the representations in $\phi(\mathcal{S})$. Constitutive assumptions or systemic laws with respect to elemental framings then correspond to upper bounds on the image $\phi(\mathcal{S})$, which we formalize in the notion of *bounding system*.

Definition 2 (Bounding systems) *A* bounding system $\mathbf{B}$ *on a set $\mathcal{D}$ consists of $\mathcal{D}$ together with a set $\mathcal{B} \subseteq \mathcal{P}(\mathcal{D})$. If $S \in \mathcal{B}$, we say S is* bounded *(with respect to* $\mathbf{B}$*).*

Thus using bounding systems $\mathbf{B}, \mathbf{B}', \ldots$ over $\mathcal{D}$ to express constitutive assumptions means we require any admissible framing $\phi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{D})$ to satisfy $\phi(\mathcal{S}) \subseteq \mathcal{B}_{\mathbf{B}} \cap \mathcal{B}_{\mathbf{B}'} \cap \ldots$. Naturally, we may collapse several bounding systems over a set to a single bounding system by taking the intersection of the several bounding sets as the combined bounding set.

We move beyond this abstract notion in three ways, each of which introduces new notions with which to describe bounding systems or constitutive assumptions with respect to elemental framings. The first way of expressing constitutional restrictions views state components as bearing an internal "logic", with the constitutionally admissible states satisfying closure and consistency conditions with respect to this logic of state components. We formalize this method for expressing constitutional restrictions using the notion of *information systems* (Scott, 1982), which builds on Tarski's theory of closure operators to abstract the structure of logical theories. The second way views at least some state components as having regulatory force, that is, as principles adopted by the agent for guiding its own reasoning, with the constitutionally admissible states satisfying the principles expressed by all the regulatory state components they contain. We formalize this in the notion of *satisfaction system* (Doyle, 1988), a clarified version of the original notion of "admissible state semantics" (Doyle, 1983c) developed to abstract traditional conceptions of deliberation and self-control. Finally, the third way of expressing constitutional restrictions views at least some state components as having motivational import, that is, as expressing the preferences, desires, or motivations of the agent in its reasoning to complement the intentional or regulatory import of satisfaction systems and logical import of information systems. We formalize this in the notion of *preference systems*, which bear some similarity to assignments of preferences to agents in political economy.

Though we employ abstract logics in describing the constitutions of agents, this serves to show the compatibility of logical notions with the theory of nonmonotonic reasoning rather than any necessity of logical notions for the theory. Indeed, the concepts of nonmonotonic reasoning make sense even for trivial abstract logics (nothing inconsistent and nothing new entailed), relying in essential ways on satisfaction and preference systems as the sources of theoretical structure.

3.2 Information systems

Many formal treatments of artificial intelligence systems employ logics to describe the beliefs of the agent by translating or rerepresenting the components of the agent's states as sentences or formulas in a logical language when the states are not already constructed as sets of logical formulae. This step of translation introduces the extra complication of a second language and the chore of translation, and, since most logics yield infinite sets of conclusions from finite sets of axioms, may prove hard to invert, making it difficult to determine the composition of the agent's state from axioms about it. We may avoid these unnecessary complications by seeking instead logics of states that capture their structure directly, that is, logics which characterize the important aspects of the structure of states that remain invariant under any rerepresentation in terms of different languages. Such logics, called *information systems* following Scott (1982), have already been developed for use in the theory of computational domains and data types.

Definition 3 (Information systems) *An* information system $\mathbf{I}$ *over a set* $\mathcal{D}$ *consists of* $\mathcal{D}$ *together with a set* $\mathcal{C} \subseteq \mathcal{P}_{\mathbf{f}}(\mathcal{D})$ *of finite subsets of* $\mathcal{D}$ *(the "consistent" finite subsets), and a relation* $\vdash$ *on* $\mathcal{C} \times \mathcal{D}$ *(the "entailment" relation), such that for each* $x, y \in \mathcal{D}$ *and* $X, Y \subseteq \mathcal{D}$*,* $\mathcal{C}$ *and* $\vdash$ *satisfy (writing* $X \vdash y$ *to mean* (X, y) *is in* $\vdash$*)*

1. *If* $X \subseteq Y \in \mathcal{C}$*, then* $X \in \mathcal{C}$*,*
2. *If* $y \in \mathcal{D}$*, then* $\{y\} \in \mathcal{C}$*,*
3. *If* $X \vdash y$*, then* $X \cup \{y\} \in \mathcal{C}$*,*
4. *If* $x \in X$*, then* $X \vdash x$*, and*
5. *If* $Y \vdash x$ *for all* $x \in X$*, and* $X \vdash y$*, then* $Y \vdash y$*.*

We say that a set $X \subseteq \mathcal{D}$ *is* consistent *iff each finite subset* $Y \subseteq X$ *is consistent according to* $\mathcal{C}$*;* (deductively) closed *iff* $x \in X$ *whenever* $Y \subseteq X$ *and* $Y \vdash x$*; and an* element *of the information system iff both closed and consistent. We write* $\mathcal{E}(\mathbf{I})$ *to denote the set of elements of* $\mathbf{I}$*. If* $\emptyset \vdash x$*, we say* x *is a* tautology *in* $\mathbf{I}$*.*

As we apply them here, each information system captures a logic of state components that we use to express closure and consistency constraints on state representations. Information systems view each element of $\mathcal{D}$ as a "proposition" about states, and each set of elements as a partial description of some state, with bigger sets representing better descriptions. The conditions on consistency and entailment state that subsets of consistent sets are consistent; that each element is itself consistent; that addition of entailed elements preserves consistency; that consistent sets entail their own members; and that entailment is transitive. This is clearer if we extend the notation of entailment in the natural way to say that $X \vdash Y$ iff $X \vdash y$ for each $y \in Y$, in which case the last condition can be rewritten as $X \vdash Z$ whenever $X \vdash Y$ and $Y \vdash Z$. Clearly $\mathcal{D}$ is closed, as are intersections of closed sets. Each information system thus determines a bounding system, namely $(\mathcal{D}, \mathcal{E}(\mathbf{I}))$, and we use such bounding systems below to require that admissible constitutional states correspond to "theories" (deductively closed and consistent sets) with respect to the constitutive information system.

Each information system gives rise to a closure operator Θ similar to the usual deductive closure operator Th of ordinary logic.

Definition 4 (Closures) *We define $\Theta(X)$, the* closure *of X, by*

$$\Theta(X) \overset{\text{def}}{=} \{x \in \mathcal{D} \mid \forall Y \subseteq \mathcal{D} \quad [(X \subseteq Y \wedge x \in X \wedge Y \vdash Y) \rightarrow x \in X]\},^{1}$$

and broaden the notation $X \vdash x$ to mean $x \in \Theta(X)$.

The closure of a set is thus its the least closed superset. Domain theory usually restricts attention to closures of consistent sets, but we find it useful to employ the unrestricted notion as well. The operator Θ exhibits properties similar to those of Th, and one easily checks that Θ is monotonic ($\Theta(A) \subseteq \Theta(B)$ whenever $A \subseteq B \subseteq \mathcal{D}$), idempotent ($\Theta(\Theta(A)) = \Theta(A)$), and identifies the closed sets as its fixed points (A is closed iff $A = \Theta(A)$).

Many notions of entailment and consistency fit within the framework of information systems. The ordinary logical notions of consistency and entailment clearly satisfy these axioms. To model agents with no ability to reject inconsistencies, we may take the minimal notion of consistency which considers all finite sets, and hence all sets, as consistent. Similarly, to model agents with no (automatic) inferential powers, we may take entailment to be simple containment ($X \vdash y$ iff $y \in X$), a minimal notion that considers all sets deductively closed. When both $\mathcal{C}$ and $\vdash$ are minimal, every set is closed and consistent, that is, $\mathcal{E}(\mathbf{I}) = \mathcal{P}(\mathcal{D})$. Alternatively, if some elements x have explicit contraries that we denote $\neg x$, one may choose $\mathcal{C}$ to capture only lack of overt inconsistency, so that $X \in \mathcal{C}$ iff X is finite and there is no $x \in X$ such that $\neg x \in X$ as well. Similarly, presuming some notion of logical structure of elements one may choose $\vdash$ to capture only propositional deduction, or only Modus Ponens ($X \vdash y$ iff either $y \in X$ or there is some $x \in \mathcal{D}$ such that $x \in X$ and $x \rightarrow y \in X$), or only ground instantiation ($X \vdash y$ iff either $y \in X$ or y is a ground instance of some $y' \in X$), or entailment in modal, relevance, or probabilistic logics, or logics of partial data structures, for which deductive closure amounts to filling in missing but implied "fields" to complete the data structure. In addition, one may always choose $\vdash$ so that some set of attitudes is present in every state as an unchangeable background to the agent's reasoning. These tautologies need not just be real logical tautologies, but instead may be substantial attitudes or "axioms".

Efficient mechanizability of a constitution (in the physical sense as well as the computational sense) often depends on the "locality" of the conditions imposed on states by constitutions, since local conditions involving a small or bounded set of state components often can be checked by a simple circuit or computational procedure with bounded effort, while global conditions on states may require unbounded effort to verify. We therefore usually seek to find constitutions that express conditions as locally as possible. We will not develop a precise theory of local and global properties or their relation to mechanizability in this paper (see, for example, (Abelson, 1978) for one such theory). We do, however, make use of the related notion of compactness already familiar in logic and topology.

[1] Added in the reprinting: The definition given in the original, $\Theta(X) \overset{\text{def}}{=} \{x \in \mathcal{D} \mid \exists Y \in \mathcal{C} \quad Y \subseteq X \wedge Y \vdash x\}$, was incorrect, and has been replaced in this version.

Definition 5 (Compact information systems) *We call an information system* compact *just in case whenever $A \vdash \{e\}$ in it, there exists a finite $C \subseteq A$ such that $C \vdash \{e\}$.*

Not all information systems are compact since logics correspond to information systems and some logics, such as dynamic logic, are not compact. Clearly, though, any information system with a trivial entailment relation ($X \vdash x$ iff $x \in X$) is compact.

3.3 Satisfaction systems

Standard constitutions for physical systems consist of the physical laws of nature, and we ordinarily view these laws as quite separate from the entities they relate. In psychological systems, however, it seems natural to employ laws of a somewhat different character, laws represented by state components themselves. That is, we interpret some state components as restrictions on possible states, as indicating bounds on the composition of the states in which the components occur, so that the states themselves set out part of their own constitution. This same point of view applies to physical systems as well, but it seems most natural for the intentional systems of psychology, especially in formulating constitutions for deliberating agents that restrict or guide their future behavior by adoption of plans or intentions regarding that behavior. Both senses fit (retrospectively, at least) within Boole's notion of "laws of thought", with state-inspecific restrictions capturing fixed legal constitutions of the agent, and state-specific restrictions capturing variable laws and amendments to the fixed constitution. To formalize interpretations of some mental components as bearing constitutive meaning, we employ the notion of *satisfaction system*.

Definition 6 (Satisfaction systems) *A* satisfaction system $\mathbf{S}$ *over $\mathcal{D}$ consists of $\mathcal{D}$ together with a meaning function $[\![\,]\!] : \mathcal{D} \to \mathcal{P}(\mathcal{P}(\mathcal{D}))$ that we extend to a function over subsets of $\mathcal{D}$ by $[\![\emptyset]\!] = \mathcal{P}(\mathcal{D})$ and $[\![X]\!] = \bigcap_{x \in X} [\![x]\!]$ for each nonempty $X \subseteq \mathcal{D}$. We say that a set $X \subseteq \mathcal{D}$ is* satisfying *or* component-admissible *iff $X \in [\![x]\!]$ for each $x \in X$, and write $\mathcal{Q}(\mathbf{S})$ for the set of all satisfying sets.*

Intuitively, the meaning $[\![x]\!]$ of an element x is the set of possible states that satisfy the constitutive intent, if any, of the element. If x has no constitutive import and places no restrictions on possible states containing it, we may take $[\![x]\!] = \mathcal{P}(\mathcal{D})$. Satisfying sets are then just sets that satisfy the constitutive import of each of the elements they contain, that is, sets $X \subseteq \mathcal{D}$ such that $X \in [\![X]\!]$. Note that the definition permits unsatisfiable elements (elements x such that $[\![x]\!] = \emptyset$) which may not appear in any satisfying state. Each satisfaction system $\mathbf{S}$ over $\mathcal{D}$ thus determines a bounding system, namely $(\mathcal{D}, \mathcal{Q}(\mathbf{S}))$.

The theory of nonmonotonic reasons elaborated below provides an important example of self-specification, interpreting a reason $A \setminus\!\!\setminus B \Vdash C$ as requiring satisfying states to contain each element of C if they contain each element of A and contain no element of B.

In practice, most interesting satisfaction systems concern only local conditions on finite portions of states, even when the states themselves may be infinite, much as many traditional logical consistency and closure conditions exhibit compactness properties. We define an analogous notion of compactness for satisfaction systems as follows.

Definition 7 (Compact satisfaction systems) *If $G \subseteq \mathcal{D}$, we write $\pi_G : \mathcal{P}(\mathcal{D}) \to \mathcal{P}(G)$ to denote both the natural projection function of subsets of $\mathcal{D}$ onto subsets of G and its lifting to sets of subsets. We say that $[\![d]\!]$ has* basis *$G \subseteq \mathcal{D}$ just in case $S \in [\![d]\!]$ iff $\pi_G(S) \in \pi_G([\![d]\!])$ for every $S \subseteq \mathcal{D}$, and that $[\![\,]\!]$ has basis G iff $[\![d]\!]$ has basis G for each $d \in \mathcal{D}$. We then say that a satisfaction system is* compact *iff for each $d \in \mathcal{D}$, whenever $[\![d]\!]$ has basis G there exists a finite $G' \subseteq G$ such that $[\![d]\!]$ has basis G'.*

Clearly, every meaning has basis $\mathcal{D}$, so every component meaning of a compact satisfaction systems has a finite basis. Not all interesting meanings have finite basis; in particular, the meaning $\mathcal{P}(\mathcal{D}) \setminus \{\emptyset\}$ expressing only nonemptiness of states does not have finite basis if $\mathcal{D}$ is infinite. However, the trivial meaning $\mathcal{P}(\mathcal{D})$ has finite (indeed, empty) basis, so the trivial satisfaction system assigning this meaning to each element is compact.

3.4 Preference systems

In addition to interpreting some elements of mental states as expressing intentional information about the agent's commitments (or its designer's), psychological theories also commonly interpret some elements as expressing evaluative information conveying the agent's (or designer's) desires and preferences (relative desires). We treat some of this information through the notion of *preference system*, which formalizes the interpretation of state components as expressing preference information.

Preference systems build on a notion of preference related to the notion standard in economics and decision theory.

Definition 8 (Preferences) *A* partial preference preorder *$\succsim$ over a set X is a reflexive ($x \succsim x$) and transitive ($x \succsim z$ whenever $x \succsim y$ and $y \succsim z$) binary relation on X. If $x \succsim y$, we say that x is* weakly preferred *to y. If x and y are weakly preferred to each other, we write $x \sim y$ and say that x and y are* indifferent*. If x is weakly preferred to y but y is not weakly preferred to x, we write $x \succ y$ and say that x is* strictly preferred *to y. If none of these cases hold, we say that x and y are* preferentially unrelated*.*

Clearly, indifference is an equivalence relation, and partitions the base set into a set of indifference classes. The identity relation is the smallest preference preorder on a set. It leaves all distinct elements preferentially unrelated, and hence has the largest number of indifference classes of elements (one for each element). The complete relation is the largest preference preorder. It makes all elements indifferent, and has the smallest number of indifference classes (just one).

Definition 9 (Preference system) *A* preference system $\mathbf{P}$ *over $\mathcal{D}$ consists of $\mathcal{D}$ together with a mapping $\succsim$ taking each $d \in \mathcal{D}$ to a partial preference preorder $\succsim_d \overset{\mathrm{def}}{=} \succsim(d)$ over $\mathcal{P}(\mathcal{D})$.*

We take constitutional preferences to have a quite different character than the constitutive intentions of satisfaction systems. Satisfaction systems capture the notion of aiming to satisfy all one's intentions at once, and define satisfying states as those states satisfying all of the intentions they contain. That is, satisfying states must express consistent sets of intentions. In contrast, we view the preferences expressed via preference systems as forms of desires, which need not be all consistent or satisfied, even in rational agents. In consequence, one cannot reasonably require that constitutionally admissible states satisfy (in some sense) all the preferences they express, and must settle for states that are *optimal* in the sense of satisfying as many preferences as possible.

Definition 10 (Optimal sets) *If $S \subseteq \mathcal{D}$ and $R \subseteq \mathcal{P}(\mathcal{D})$, we say that S is* (Pareto) optimal *in the range R with respect to a preference system $\succsim$ over $\mathcal{D}$ just in case whenever we have $S' \succ_d S$ for some $S' \in R$ and $d \in S'$, we also have $S \succ_e S'$ for some $e \in S$.*

That is, S is optimal in R if every preference for some set in R expressed by the other set can be countered by a preference for S expressed by S. Clearly, if $R' \subseteq R$ and S is optimal in R, then S is also optimal in R'. Thus if S is optimal in $\mathcal{P}(\mathcal{D})$, it is optimal in every range.

We do not identify any standard constitutional restriction based on optimality due to the variability of range within one judges optimality, and instead consider optimality with respect to specific classes of preferences and specific ranges. We hope further investigation either identifies natural choices of range for constitutional restrictions, or explains why such choices do not exist.

3.5 Basic constitutions

Our basic notion of constitution combines bounds expressed by satisfaction and information systems with general bounds.

Definition 11 (Basic constitution) *A* basic constitution Σ *over* $\mathcal{D}$ *consists of an information system* $\mathbf{I} = (\mathcal{D}, \vdash, \mathcal{C})$*, satisfaction system* $\mathbf{S} = (\mathcal{D}, [\![\,]\!])$*, and bounding system* $\mathbf{B} = (\mathcal{D}, \mathcal{B})$ *over* $\mathcal{D}$*, with the* admissible states $\mathcal{A}(\Sigma)$ *of the constitution being the states jointly characterized by* $\mathcal{E}(\mathbf{I})$*,* $\mathcal{Q}(\mathbf{S})$*, and* $\mathcal{B}(\mathbf{B})$*, that is, the sets* $S \subseteq \mathcal{D}$ *such that* $\mathcal{P}_{\mathbf{f}}(S) \subseteq \mathcal{C}$*,* $S = \Theta(S)$*,* $S \in [\![S]\!]$*, and* $S \in \mathcal{B}$*.*

The notion of basic constitution recasts the notion of admissible state semantics of (Doyle, 1983a, 1983c) in simpler terms. The treatment in (Doyle, 1983c) assigned constitutive intentional meaning to elements and judged states satisfying or not just as in the present treatment, and also provided what we now call a bounding system through an arbitrary restriction set $\mathcal{R} \subseteq \mathcal{P}(\mathcal{D})$. The main difference in formulation concerned the background logic, which was not distinguished explicitly in an information system. The specific theories developed in (Doyle, 1983c) instead defined restriction sets consisting of the sets closed with respect to a compact closure operator, ignoring any notion of consistency not expressed by the satisfaction system meanings.

In many cases, we do not need the full generality of basic constitutions, and characterize states in terms of only one or two of the information, satisfaction, or bounding systems of a basic constitution. We give names to these special constitutions as follows, summarizing the nomenclature in Table 1.

Definition 12 (Special constitutions) *We call a basic constitution over* $\mathcal{D}$ logical *just in case its information system characterizes its admissible states;* satisfying *just in case its satisfaction system characterizes its admissible states;* logically satisfying *just in case its information and satisfaction systems jointly characterize its admissible states;* unstructured *just in case its bounding system characterizes its admissible states, and* trivial *just in case all subsets of* $\mathcal{D}$ *are admissible states. We call a state space* $\mathcal{S}$ logical, satisfying, logically satisfying, *or* trivial *with respect to an elemental framing* ϕ *over* $\mathcal{D}$ *if* $\phi(\mathcal{S})$ *can be characterized by a logical, satisfying, logically satisfying, or trivial constitution over* $\mathcal{D}$*.*

Satisfying constitutions, in which the states themselves explicitly express all restrictions on states, have a certain philosophical attraction, and recall efforts in artificial intelligence towards constructing completely "self-descriptive" machines (e.g., (de Kleer et al., 1977; Doyle, 1980; McDermott, 1978; Minsky, 1965)).

Although the different parts of basic constitutions sometimes permit one to shift restrictions from one subsystem to another, different subtypes of constitutions have different expressive powers. We begin by characterizing the simplest constitutions, the trivial ones.

Theorem 13 (Trivial constitutions) *A basic constitution* Σ *is trivial iff*

if $\mathcal{A}(\Sigma)$ *is*	*then* Σ *is*
$\mathcal{E}(\mathbf{I}(\Sigma))$	logical
$\mathcal{Q}(\mathbf{S}(\Sigma))$	satisfying
$\mathcal{E}(\mathbf{I}(\Sigma)) \cap \mathcal{Q}(\mathbf{S}(\Sigma))$	logically satisfying
$\mathcal{B}(\mathbf{B}(\Sigma))$	unstructured
$\mathcal{P}(\mathcal{D})$	trivial

Table 1: Special constitutions

1. $\mathcal{B}(\Sigma) = \mathcal{P}(\mathcal{D})$,
2. $\mathcal{C} = \mathcal{P}_{\mathbf{f}}(\mathcal{D})$ *and* $\vdash$ *is* $\ni$ *(i.e.,* $\in^{-1}$*), and*
3. $\{X \subseteq \mathcal{D} \mid d \in X\} \subseteq [\![d]\!]$ *for every* $d \in \mathcal{D}$.

PROOF: Suppose Σ is a basic constitution over $\mathcal{D}$.

(If) Suppose conditions (1)–(3) hold. The first part of condition (2) implies that all sets are consistent, while the second part implies that all sets are closed, hence $\mathcal{E}(\mathbf{I}(\Sigma)) = \mathcal{P}(\mathcal{D})$. Furthermore, condition (3) implies that $x \in [\![d]\!]$ for each $d \in X \subseteq \mathcal{D}$, so every $X \subseteq \mathcal{D}$ is satisfying, hence $\mathcal{Q}(\mathbf{S}(\Sigma)) = \mathcal{P}(\mathcal{D})$. Thus $\mathcal{A}(\Sigma) = \mathcal{P}(\mathcal{D})$.

(Only if) Suppose $\mathcal{A}(\Sigma) = \mathcal{P}(\mathcal{D})$. Since $\mathcal{A}(\Sigma) = \mathcal{B}(\Sigma) \cap \mathcal{E}(\mathbf{I}(\Sigma)) \cap \mathcal{Q}(\mathbf{S}(\Sigma))$, we clearly have $\mathcal{B}(\Sigma) = \mathcal{P}(\mathcal{D})$, which is condition (1), as well as $\mathcal{E}(\mathbf{I}(\Sigma)) = \mathcal{Q}(\mathbf{S}(\Sigma)) = \mathcal{P}(\mathcal{D})$. Since every set is consistent, so is every finite subset of $\mathcal{D}$, so we have $\mathcal{C} = \mathcal{P}_{\mathbf{f}}(\mathcal{D})$. Next, if we have $X \subseteq \mathcal{D}$ and $X \vdash e$ for some $e \notin X$, then X is not closed, hence $X \notin \mathcal{E}(\mathbf{I}(\Sigma))$. Since every X is closed, this means $\vdash$ is $\ni$. Finally, if $d \in X \notin [\![d]\!]$, then X is not satisfying, hence $\mathcal{Q}(\mathbf{S}(\Sigma)) \neq \mathcal{P}(\mathcal{D})$. But every X is satisfying, so $X \in [\![d]\!]$ whenever $d \in X$, hence $\{X \subseteq \mathcal{D} \mid d \in X\} \subseteq [\![d]\!]$. □

We may characterize the expressive power of satisfying constitutions quite easily.

Theorem 14 (Satisfying constitutions) *A state space* $\mathcal{S}$ *is satisfying with respect to an elemental framing* ϕ *over* $\mathcal{D}$ *iff* $\emptyset \in \phi(\mathcal{S})$.

PROOF: Let ϕ be an elemental framing of $\mathcal{S}$ over $\mathcal{D}$. If $\emptyset \in \phi(\mathcal{S})$, define $[\![\;]\!]$ over $\mathcal{D}$ so that $[\![d]\!] = \phi(\mathcal{S})$ for every $d \in \mathcal{D}$. Then nonempty subsets of $\mathcal{D}$ are satisfying iff they are in $\phi(\mathcal{S})$, and $\emptyset$ is both satisfying and in $\phi(\mathcal{S})$, hence $\phi(\mathcal{S})$ is exactly the set of satisfying states, so $\mathcal{S}$ is satisfying with respect to ϕ. Conversely, if $\mathcal{S}$ is satisfying with respect to ϕ, then $\emptyset \in \phi(\mathcal{S})$ since $\emptyset$ is always satisfying. □

Thus satisfaction systems alone cannot characterize all interesting elemental framings, since they cannot express bounding systems that require nonemptiness of state representations. More generally, even if $\emptyset \notin \phi(\mathcal{S})$, the meaning function constructed in the proof of Theorem 14 yields $\phi(\mathcal{S}) \cup \{\emptyset\}$ as the set of satisfying sets (though this encoding makes all restrictions global and may convert meanings of finite basis into meanings of infinite basis since $\phi(\mathcal{S})$ also need not have finite basis). In consequence, any state space may be characterized by a satisfaction system together with the bounding system consisting of all nonempty sets of state components, that is $(\mathcal{D}, \mathcal{P}(\mathcal{D}) \setminus \{\emptyset\})$. Unfortunately, one cannot always achieve the same effect using logical restrictions instead of general bounds.

Theorem 15 (Nonemptiness) *If $\phi(\mathcal{S}) = \mathcal{P}(\mathcal{D}) \setminus \{\emptyset\}$ and $\mathcal{D}$ has at least two elements, then $\phi(\mathcal{S})$ is not logically satisfying.*

PROOF: Suppose $\phi(\mathcal{S})$ is jointly characterized by a satisfaction and information system, $\mathcal{D}$ has at least two elements, and $\emptyset \notin \phi(\mathcal{S})$. Since $\emptyset$ is always satisfying, we must have $\emptyset \neq \Theta(\emptyset)$. Now the set of tautologies $\Theta(\emptyset)$ either contains at least two state components or does not. If it contains at least two, then no set $\{d\}$ consisting of only one tautology is closed, hence $\phi(\mathcal{S}) \neq \mathcal{P}(\mathcal{D}) \setminus \{\emptyset\}$. On the other hand, if there is just one tautology $\{d\}$, then no $\{d'\}$ is not closed when $d \neq d'$, so again $\phi(\mathcal{S}) \neq \mathcal{P}(\mathcal{D}) \setminus \{\emptyset\}$. □

Thus basic constitutions for some state spaces must have nontrivial bounding systems expressing at least the condition of nonemptiness, that is, $(\mathcal{D}, \mathcal{P}(\mathcal{D}) \setminus \{\emptyset\})$.

We have no complete characterization of how logical and satisfying constitutions differ in expressive power. One observes immediately that some logical state spaces are not satisfying, since the empty set is always satisfying, but if a state space is characterized by an information system containing tautologies, then $\emptyset \notin \phi(\mathcal{S})$. Conversely, some satisfying state spaces are not logical, since a satisfaction system may assign $[\![d]\!] = \emptyset$, in which case no state contains d, even though we must have $\{d\} \in \mathcal{C}$ by the axioms for information systems. We also observe without proof that if $[\![d]\!] \neq \emptyset$ for every $d \in \mathcal{D}$, then one can construct an information system from $[\![\;]\!]$ by defining $\mathcal{C}$ to contain those finite sets $X \subseteq \mathcal{D}$ such that $[\![X]\!] \neq \emptyset$, and by defining $X \vdash x$ to hold just in case $[\![X]\!] \subseteq [\![x]\!]$. Unfortunately, the elements of this constructed information system have no particular relation to the satisfying sets of the original satisfaction system. While satisfying states must be consistent in the constructed information system, they need not be closed, and conversely, consistent closed sets need not be satisfying.

3.6 Preferential constitutions

We obtain a larger conception of constitution by augmenting basic constitutions with preference systems to obtain *preferential* constitutions.

Definition 16 (Preferential constitutions) *A* preferential *basic constitution combines a basic constitution Σ over $\mathcal{D}$ with a preference system $\succsim$ over $\mathcal{D}$.*

Unlike the way in which the information, satisfaction, and bounding systems of basic constitutions restrict the states admissible with respect to the constitution, we do not require that the preference system of a preferential basic constitution yields any restrictions on admissible states, though specific preferential basic constitutions may in fact restrict admissible states to those satisfying certain preferential conditions. The most natural candidates for such preferential restrictions are optimality with respect to specific classes of preferences and ranges. Accordingly, we define optimality of constitutions with respect to specific classes of preferences by considering whether all admissible states are optimal among admissible states.

Definition 17 (Constitutional optimality) *A basic constitution Σ over $\mathcal{D}$ is* optimal *with respect to a preference system $\succsim$ over $\mathcal{D}$ just in case each $S \in \mathcal{A}(\Sigma)$ is optimal in $\mathcal{A}(\Sigma)$ with respect to $\succsim$.*

Clearly, one can always find constitutions optimal with respect to any preference system, since if $\succsim$ is a preference system over $\mathcal{D}$, the unstructured basic constitution Σ over $\mathcal{D}$ such that $\mathcal{B}(\Sigma) = \{S\}$ for some $S \subseteq \mathcal{D}$ is trivially optimal.

Preference systems preferring satisfying states to unsatisfying states provide natural connections between satisfying constitutions and optimal constitutions.

Definition 18 (Satisfaction preferences) *The* pure satisfaction preference *system corresponding to a satisfaction system* $(\mathcal{D}, [\![\,]\!])$ *is the preference system* $(\mathcal{D}, \succsim)$ *such that for each* $d \in \mathcal{D}$ *and* $S, S' \subseteq \mathcal{D}$*, we have* $S \succ_d S'$ *just in case* $S \in [\![d]\!]$ *and* $S' \notin [\![d]\!]$*, and* $S \sim_d S'$ *just in case* $S \in [\![d]\!]$ *iff* $S' \in [\![d]\!]$*. The* weak present satisfaction *preference system defines* $S \succ_d S'$ *just in case* $d \in S \in [\![d]\!]$ *but not* $d \in S' \in [\![d]\!]$ *(that is, either* $d \notin S'$ *or* $S' \notin [\![d]\!]$*), while the* strong present satisfaction *preference system defines* $S \succ_d S'$ *just in case* $d \in S \in [\![d]\!]$ *but* $d \in S' \notin [\![d]\!]$*. We call sets optimal in some range with respect to the (weak, strong present) satisfaction preference system* (weakly, strongly present) satisfaction optimal *in the range.*

We then have the following result.

Theorem 19 (Satisfaction optimality) *Every satisfying constitution is optimal with respect to the corresponding strong present satisfaction preference system.*

PROOF: Let Σ be a satisfying constitution over $\mathcal{D}$, and let $\succsim$ be the strong present satisfaction preference system over $\mathcal{D}$ corresponding to the satisfaction system of Σ. Strong present satisfaction optimality of S means that whenever $d \in S' \in [\![d]\!]$ and $d \in S \notin [\![d]\!]$ for $S' \in \mathcal{Q}(\Sigma)$, then for some $e \in S \in [\![e]\!]$ we have $e \in S' \notin [\![e]\!]$. But since S is satisfying, $d \in S \notin [\![d]\!]$ never obtains, so S is trivially optimal among $\mathcal{Q}(\Sigma)$. □

Thus every satisfying state is strongly present satisfaction optimal among satisfying states. The converse, however, does not hold in general. For example, if $\mathcal{D} = \{d\}$ and $[\![d]\!] = \{\emptyset\}$, then $\emptyset$ is the only satisfying state, and $\mathcal{D}$ is trivially strongly present satisfaction optimal in $\{\emptyset\}$ since there is no $x \in \emptyset \in [\![x]\!]$.

4. Reasons

We may use the notions of satisfaction and preference systems to identify the entities we call *reasons*, which cover all or part of familiar notions of monotonic and nonmonotonic justifications of reason maintenance systems, default rules of default logic, and modal default rules of nonmonotonic logics, as well as some notions of taxonomic inference and virtual context mechanisms. We identify reasons by their constitutive import, in terms of the role they play in the agent's constitution, rather than in terms of mere overt structural or syntactic appearances, which vary far too much to accommodate other concerns to serve as useful guides to understanding. That is, reasons need not reflect any overt structures in the unframed states themselves. They instead form part of the metalanguage in which we discuss agent states. The agent's "language of thought", if any, need not express reasons in the same way, and we may thus find reasons in neural networks, business organizations, or even literary works, as well as in traditional sentential or computational systems representing them as straightforward data-structures that directly suggest their characteristic interpretation.

We will not develop the theory of reasons here in its full generality. Instead we focus attention on *simple* reasons, which we denote by expressions of the form $A \setminus\!\!\setminus B \mathrel{|\!\!\!-} C$, read "$A$ without B gives C", meaning roughly that if the "antecedent" items in A are present in a state but the "qualification" items in B are not, then the "conclusion" items in C should be present as well. One can extend much of the development below to cover more general reasons that also specify a set of conditionally "denied" items D, notated $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D$, but we generally avoid that additional complexity here.

4.1 Ranges and range conditionals

Our analytical language expresses reasons in terms of the notions of *propositional ranges* and *range conditionals*. Propositional ranges provide a simple way of identifying a restricted range of sets of state components to consider, namely those sets that lie between a "lower bound" set and an "upper bound" set (expressed in terms of its complement).

Definition 20 (Propositional ranges) *A* propositional range *over $\mathcal{D}$ is a set $R \subseteq \mathcal{P}(\mathcal{D})$ such that for some $A, B \subseteq \mathcal{D}$ and every $S \subseteq \mathcal{D}$ we have $S \in R$ iff $A \subseteq S$ and $S \cap B = \emptyset$ (alternatively written $S \subseteq \overline{B}$, where $\overline{B} \overset{\text{def}}{=} \mathcal{D} \setminus B$), in which case we write $A \mathbin{\backslash\!\backslash} B$ (read "A without B") to denote the range, that is,*

$$A \mathbin{\backslash\!\backslash} B \overset{\text{def}}{=} \{S \subseteq \mathcal{D} \mid A \subseteq S \subseteq \overline{B}\}.$$

We call an element $d \in \mathcal{D}$ a propositional range *with respect to a satisfaction system $(\mathcal{D}, [\![\,]\!])$ iff $[\![d]\!]$ is a propositional range over $\mathcal{D}$.*

Algebraically, of course, a range consists of the full sublattice of $\mathcal{P}(\mathcal{D})$ between two points of $\mathcal{P}(\mathcal{D})$, which we may think of as the set $\{A \cup S \mid S \subseteq \mathcal{D} \setminus (A \cup B)\}$. In particular, $\emptyset \mathbin{\backslash\!\backslash} \emptyset$ represents the complete range $\mathcal{P}(\mathcal{D})$.

We first observe that expressions like $A \mathbin{\backslash\!\backslash} B$ characterize ranges pretty well. Empty ranges admit some variability of expression, but a variability easily recognized, as seen in the following theorem.

Theorem 21 (Empty ranges) *A range $A \mathbin{\backslash\!\backslash} B$ is empty iff A and B intersect.*

PROOF: Clearly, if $A \cap B \neq \emptyset$, then $A \mathbin{\backslash\!\backslash} B$ is empty. Conversely, if $A \mathbin{\backslash\!\backslash} B = \emptyset$, then by definition for each $S \subseteq \mathcal{D}$ we have either $A \not\subseteq S$ or $S \cap B \neq \emptyset$. If we choose $S = A$, then $A \subseteq S$, so we have $A \cap B \neq \emptyset$ as desired. □

Nonempty ranges, in contrast, have exactly one expression of the form $A \mathbin{\backslash\!\backslash} B$.

Theorem 22 (Unique representation) *If $A \mathbin{\backslash\!\backslash} B = A' \mathbin{\backslash\!\backslash} B' \neq \emptyset$, then $A = A'$ and $B = B'$.*

PROOF: Suppose $A \mathbin{\backslash\!\backslash} B = A' \mathbin{\backslash\!\backslash} B' \neq \emptyset$. By Theorem 21 we have $A \cap B = A' \cap B' = \emptyset$. Thus $A \in A \mathbin{\backslash\!\backslash} B$, so $A \in A' \mathbin{\backslash\!\backslash} B'$, so $A' \subseteq A$. But we also have $A' \in A' \mathbin{\backslash\!\backslash} B'$, so $A' \in A \mathbin{\backslash\!\backslash} B$, so $A \subseteq A'$, hence $A = A'$. We next observe that $B \setminus B'$ must be empty, since if $x \in B \setminus B'$, we would have $A \cup \{x\} \in A' \mathbin{\backslash\!\backslash} B'$ but $A \cup \{x\} \notin A \mathbin{\backslash\!\backslash} B$, contrary to hypothesis. A similar argument shows that $B' \setminus B$ must be empty, so $B = B'$. □

To summarize, all propositional ranges have unique representations in terms of disjoint lower and upper bounds, except for the empty range, for which any overlapping bounds provide a representation.

In addition to ranges themselves, our analysis also makes heavy use of conditionals formed from ranges, which we call *range conditionals*.

Definition 23 (Range conditionals) *A* range conditional *is a set $R \subseteq \mathcal{P}(\mathcal{D})$ such that for some $A, B, C, D \subseteq \mathcal{D}$, we have $S \in R$ iff either $S \notin A \mathbin{\backslash\!\backslash} B$ or $S \in C \mathbin{\backslash\!\backslash} D$, in which case we write $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D$ (read "A without B gives C without D") to denote the range conditional, that is,*

$$A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D \overset{\text{def}}{=} \{S \subseteq \mathcal{D} \mid S \in A \mathbin{\backslash\!\backslash} B \to S \in C \mathbin{\backslash\!\backslash} D\}.$$

If $R = A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} \emptyset$ for some A, B, C, we call the range conditional simple *and abbreviate it $A \mathbin{\backslash\!\backslash} B \Vdash C$. We call an element $d \in \mathcal{D}$ a* (simple) range conditional *with respect to a satisfaction system $(\mathcal{D}, [\![\,]\!])$ iff $[\![d]\!]$ is a (simple) range conditional over $\mathcal{D}$.*

That is, range conditionals stipulate that the "conclusions" C must be held and the "denials" D must not be held if the "antecedents" A are held and none of the "qualifiers" B are held. Putting together the definitions of propositional ranges and range conditionals, we see that the range conditional $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D$ means $\{S \subseteq \mathcal{D} \mid A \subseteq S \subseteq \overline{B} \to C \subseteq S \subseteq \overline{D}\}$, and the simple range conditional $A \mathbin{\backslash\!\backslash} B \Vdash C$ means $\{S \subseteq \mathcal{D} \mid A \subseteq S \subseteq \overline{B} \to C \subseteq S\}$. We also observe that if $A = B = \emptyset$, the range conditional $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D$ reduces to the range $C \mathbin{\backslash\!\backslash} D$, that is $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D = C \mathbin{\backslash\!\backslash} D$.

In contrast to the unique representations of nonempty ranges, different range conditional expressions may mean the same thing. The most obvious variability is that antecedents (qualifications) may be added to or subtracted from the conclusions (denials) without changing the meaning.

Theorem 24 (Range reduction) *For each $A, B, C, D \subseteq \mathcal{D}$, $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D = A \mathbin{\backslash\!\backslash} B \Vdash (C \setminus A) \mathbin{\backslash\!\backslash} (D \setminus B)$.*

PROOF: Suppose $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D = A \mathbin{\backslash\!\backslash} B \Vdash (C \setminus A) \mathbin{\backslash\!\backslash} (D \setminus B)$ for $A, B, C, D \subseteq \mathcal{D}$. If $S \notin A \mathbin{\backslash\!\backslash} B$, then S is in both conditional ranges, so suppose $S \in A \mathbin{\backslash\!\backslash} B$, that is, $A \subseteq S \subseteq \overline{B}$. If $S \in A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D$, then $S \in C \mathbin{\backslash\!\backslash} D$, hence we have $C \setminus A \subseteq C \subseteq S \subseteq \overline{D} \subseteq \overline{D \setminus B}$, so $S \in A \mathbin{\backslash\!\backslash} B \Vdash (C \setminus A) \mathbin{\backslash\!\backslash} (D \setminus B)$. And if $S \in A \mathbin{\backslash\!\backslash} B \Vdash (C \setminus A) \mathbin{\backslash\!\backslash} (D \setminus B)$, then $C \setminus A \subseteq S \subseteq \overline{D \setminus B}$. But since $A \subseteq S \subseteq \overline{B}$, we have $A \cup C \subseteq S$ and $S \subseteq \overline{B \cup D}$, hence $C \subseteq S \subseteq \overline{D}$, so $S \in A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D$. Thus $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D = A \mathbin{\backslash\!\backslash} B \Vdash (C \setminus A) \mathbin{\backslash\!\backslash} (D \setminus B)$. □

Accordingly, we call a range conditional expression $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D$ *reduced* just in case neither A and C nor B and D overlap, that is, $C \setminus A = C$ and $D \setminus B = D$. To avoid the complications introduced by this variability, we ordinarily assume reduced expressions in the following.

As with trivial ranges, we may characterize trivial simple range conditionals in a simple way.

Theorem 25 (Trivial conditionals) *For each $A, B, C, D \subseteq \mathcal{D}$, we have $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D = \mathcal{P}(\mathcal{D})$ iff either $A \cap B \neq \emptyset$ or $C \setminus A = D \setminus B = \emptyset$.*

PROOF: Since $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D = \{S \subseteq \mathcal{D} \mid A \subseteq S \subseteq \overline{B} \to C \subseteq S \subseteq \overline{D}\}$, it is clear that $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D = \mathcal{P}(\mathcal{D})$ if either $A \cap B \neq \emptyset$ or $C \setminus A = D \setminus B = \emptyset$ hold. So suppose $A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D = \mathcal{P}(\mathcal{D})$ and that $C \setminus A \neq \emptyset$ or $D \setminus B \neq \emptyset$. If $C \setminus A \neq \emptyset$, then $C \not\subseteq A$, and since $A \in A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D$, we know that $A \notin A \mathbin{\backslash\!\backslash} B$. But this can happen only if $A \mathbin{\backslash\!\backslash} B = \emptyset$, which by Theorem 21 occurs only if $A \cap B \neq \emptyset$. On the other hand, if $D \setminus B \neq \emptyset$, then $D \not\subseteq B$, so $\overline{B} \not\subseteq \overline{D}$, and since $\overline{B} \in A \mathbin{\backslash\!\backslash} B \Vdash C \mathbin{\backslash\!\backslash} D$, we know that $\overline{B} \notin A \mathbin{\backslash\!\backslash} B$. Again, this can happen only if $A \mathbin{\backslash\!\backslash} B = \emptyset$, which by Theorem 21 occurs only if $A \cap B \neq \emptyset$. Thus in either case we have $A \cap B \neq \emptyset$. □

In particular, this means that a reduced simple range conditional $A \mathbin{\backslash\backslash} B \Vdash C$ is trivial iff either $A \cap B = \emptyset$ or $C = \emptyset$.

Unlike nontrivial ranges, which have unique representations, some nontrivial simple range conditionals have different representations: for example, $\emptyset \mathbin{\backslash\backslash} \{a, b\} \Vdash \{c\} = \emptyset \mathbin{\backslash\backslash} \{a, c\} \Vdash \{b\} = \emptyset \mathbin{\backslash\backslash} \{b, c\} \Vdash \{a\}$. However, the following characterization of nontrivial simple range conditionals shows that such ambiguity occurs only when some reduced representation has just one consequent and some qualifiers, as in this example, and that the antecedents of any alternative representations must coincide in any case.

Theorem 26 (Nontrivial conditionals) *If $A \mathbin{\backslash\backslash} B \Vdash C = A' \mathbin{\backslash\backslash} B' \Vdash C' \neq \mathcal{P}(\mathcal{D})$, A, B, and C are mutually disjoint, and A', B', and C' are mutually disjoint, then $A = A'$ and either (1) $B = B'$ and $C = C'$, or (2) $|C| = |C'| = 1$, $B \setminus B' = C'$, and $B' \setminus B = C$.*

PROOF: Suppose $A \mathbin{\backslash\backslash} B \Vdash C = A' \mathbin{\backslash\backslash} B' \Vdash C' \neq \mathcal{P}(\mathcal{D})$, A, B, and C are mutually disjoint, and A', B', and C' are mutually disjoint.

By nontriviality, we know that neither C nor C' is empty. Thus $A \notin A \mathbin{\backslash\backslash} B \Vdash C$, so $A \notin A' \mathbin{\backslash\backslash} B' \Vdash C'$, hence $A' \subseteq A$. Similarly, $A' \notin A' \mathbin{\backslash\backslash} B' \Vdash C'$, so $A' \notin A \mathbin{\backslash\backslash} B \Vdash C$, hence $A \subseteq A'$. Thus $A = A'$.

We also observe that if $|C| > 1$, then $C = C'$. To see this, suppose $|C| > 1$. If $x \in C$, then since C has other elements, we have $A \cup \{x\} \notin A \mathbin{\backslash\backslash} B \Vdash C$, so $A \cup \{x\} \notin A' \mathbin{\backslash\backslash} B' \Vdash C'$, and thus $x \notin B'$. Since this holds for all x in C, we have $C \cap B' = \emptyset$. Since $A \cup C \in A \mathbin{\backslash\backslash} B \Vdash C$, so $A \cup C \in A' \mathbin{\backslash\backslash} B' \Vdash C'$. Since $C \cap B' = \emptyset$, this means $C' \subseteq A \cup C$, and since A and C' are disjoint, this means $C' \subseteq C$. But we also have $A \cup C' \in A' \mathbin{\backslash\backslash} B' \Vdash C'$, so $A \cup C' \in A \mathbin{\backslash\backslash} B \Vdash C$ as well. But $A \cup C' \in A \mathbin{\backslash\backslash} B$, since $C' \subseteq C$ and $C \cap B' = \emptyset$. Thus $C \subseteq A \cup C'$, so $C \subseteq C'$, so $C = C'$.

We now prove the assertion by cases.

If $C = C'$, suppose by way of contradiction that $B \setminus B' \neq \emptyset$, and let $z \in B \setminus B'$. Then $A \cup \{z\} \in A \mathbin{\backslash\backslash} B \Vdash C$, so $A \cup \{z\} \in A' \mathbin{\backslash\backslash} B' \Vdash C'$. Since $z \in B$, the disjointness and equality assumptions imply $z \notin C'$, so $C' \not\subseteq A \cup \{z\}$, thus $A \cup \{z\} \notin A' \mathbin{\backslash\backslash} B'$, hence $z \in B'$, contrary to hypothesis. A symmetric argument shows that we cannot have $B' \setminus B \neq \emptyset$, so we conclude $B = B'$.

If $C \neq C'$, we must have $|C| = |C'| = 1$, since $|C| > 1$ or $|C'| > 1$ imply that $C = C'$. Now if $z \in B \setminus B'$, then $A \cup \{z\} \in A \mathbin{\backslash\backslash} B \Vdash C$, so $A \cup \{z\} \in A' \mathbin{\backslash\backslash} B' \Vdash C'$, and since $A \cup \{z\} \in A' \mathbin{\backslash\backslash} B'$, we have $C' \subseteq A \cup \{z\}$, which by the disjointness assumptions means $C' = \{z\}$. Since this holds for every such z, we see that $B \setminus B' = C'$. A similar argument about elements of $B' \setminus B$ shows that $B' \setminus B = C$.

Summarizing, if $C = C'$, then $B = B'$ as well, while if $C \neq C'$, we have $|C| = |C'| = 1$, $B \setminus B' = C'$, and $B' \setminus B = C$. These constitute the two cases claimed in the statement of the theorem. □

While simple range conditionals prove important in the theory and practice of artificial intelligence, they cannot express all interesting state specifications. In particular, they cannot exclude any state component from admissible states.

Theorem 27 (Inclusivity) *If $\mathcal{D}$ contains only simple range conditionals, then $\mathcal{D}$ is satisfying.*

PROOF: If $\mathcal{D}$ contains only simple range conditionals and $d \in \mathcal{D}$ with $\llbracket d \rrbracket = A \mathbin{\backslash\backslash} B \Vdash C$, then $A \subseteq \mathcal{D} \subseteq \overline{B} \rightarrow C \subseteq \mathcal{D}$ is trivially true, so $\mathcal{D} \in \llbracket d \rrbracket$. Since this holds for every $d \in \mathcal{D}$, $\mathcal{D}$ is satisfying. □

Expressing more state spaces thus requires state components that do not represent simple range conditionals (even trivial ones).

We earlier saw that the strongly present satisfaction optimal sets include the satisfying states. Simple range conditionals, however, yield an exact correspondence between these notions.

Theorem 28 (Conditional satisfaction optimality) *If $\mathcal{D}$ contains only simple range conditionals and $S \subseteq \mathcal{D}$ is strongly present satisfaction optimal in $\mathcal{P}(\mathcal{D})$, then S is satisfying.*

PROOF: We prove the contrapositive. Suppose $S \subseteq \mathcal{D}$ is unsatisfying, that is, for some $d \in \mathcal{D}$, $d \in S \notin [\![d]\!]$. Now by Theorem 27, $\mathcal{D} \in [\![d]\!]$, but for all $e \in S$, $\mathcal{D} \in [\![e]\!]$ as well, so S is not satisfaction optimal. □

In light of Theorem 19, we conclude that when all state components are simple range conditionals, sets are satisfying iff they are strongly present satisfaction optimal.

4.2 Simple reasons

We now define *simple reasons* in terms of range conditionals and preference systems.

Definition 29 (Simple Reasons) *An element $d \in \mathcal{D}$ is a (finite)* simple reason *with respect to a satisfaction system $(\mathcal{D}, [\![\,]\!])$ and preference system $(\mathcal{D}, \succsim)$ iff there are (finite) disjoint sets $A, B, C \subseteq \mathcal{D}$ such that*

1. *$[\![d]\!] = A \setminus\!\!\setminus B \Vdash C$, and*

2. *For each $S, S' \in [\![d]\!]$ such that $A \subseteq S, S'$, if $S \subseteq \overline{B}$ and $S' \not\subseteq \overline{B}$, then $S \succ_d S'$.*

We call each triple of sets A, B, C satisfying these conditions a simple reason interpretation *of d, and (abusing the notation slightly) write $A \setminus\!\!\setminus B \Vdash C$ to indicate their order. We say that a simple reason d is* trivial *if $[\![d]\!] = \mathcal{P}(\mathcal{D})$, and that it is* nontrivial *otherwise.*

Thus a simple reason indicates both a range conditional stipulating that the "conclusions" C must be held if the "antecedents" A are held and none of the "qualifiers" B are held, and a preference for not holding any qualifiers if the antecedents are held. The preference order required of a reason in this definition is both partial and considerably weaker than the total preference order assigned to the corresponding default rule in (Doyle & Wellman, 1991). The present development does not require us to stipulate a total order since the range conditional rules out some of the states compared by the total preference order. The two interpretations agree, however, in that the the total order contains the preference comparisons needed to identify simple reasons.

The constitutive preference employed in identifying reasons provides the simple range conditional with a canonical form, which we see as follows. The first step towards this end comes directly from our analysis of trivial range conditionals, which we use to characterize trivial reasons as ones placing no restrictions on states.

Theorem 30 (Trivial reasons) $[\![d]\!] = \mathcal{P}(\mathcal{D})$ *iff d is a simple reason with reason interpretation $\emptyset \setminus\!\!\setminus \emptyset \Vdash \emptyset$.*

PROOF: Clearly, if d is a simple reason with reason interpretation $\emptyset \setminus\!\!\setminus \emptyset \mid\!\vdash \emptyset$, then $[\![d]\!] = \mathcal{P}(\mathcal{D})$, so suppose $[\![d]\!] = \mathcal{P}(\mathcal{D})$. Then we clearly have $[\![d]\!] = \emptyset \setminus\!\!\setminus \emptyset \mid\!\vdash \emptyset$. Furthermore, d vacuously satisfies the second part of the definition of simple reason no matter what preference order it indicates since there is no S' with $S' \not\subseteq \overline{\emptyset} = \mathcal{D}$. Thus d can be interpreted as the simple reason $\emptyset \setminus\!\!\setminus \emptyset \mid\!\vdash \emptyset$. □

Note well that different trivial reasons may bear different preferential interpretations, for as the proof of Theorem 30 noted, their preference interpretations cannot alter their interpretation as reasons.

For nontrivial reasons, we have a stronger characterization.

Theorem 31 (Nontrivial reasons) *If d is a nontrivial simple reason, it has a unique simple reason interpretation.*

PROOF: Suppose $[\![d]\!] = A \setminus\!\!\setminus B \mid\!\vdash C = A' \setminus\!\!\setminus B' \mid\!\vdash C' \neq \mathcal{P}(\mathcal{D})$, A, B, and C are mutually disjoint, A', B', and C' are mutually disjoint, and for all $S, S' \subseteq \mathcal{D}$ we have

1. If $A \subseteq S, S'$, $S \subseteq \overline{B}$, and $S' \not\subseteq \overline{B}$, then $S \succ_d S'$, and
2. If $A' \subseteq S, S'$, $S \subseteq \overline{B'}$, and $S' \not\subseteq \overline{B'}$, then $S \succ_d S'$.

We first prove $C = C'$. Suppose, by way of contradiction, that $C \neq C'$. Theorem 26 then indicates that $B \setminus B'$ and $B' \setminus B$ are nonempty, so let $S = A \cup (B \setminus B')$ and $S' = A \cup (B' \setminus B)$. Clearly, $S \neq S'$, and we have $S \subseteq \overline{B'}$ and $S \not\subseteq \overline{B}$ and $S' \subseteq \overline{B}$ and $S' \not\subseteq \overline{B'}$. Applying the two enumerated hypotheses above, this yields $S \succ_d S' \succ_d S$, contradicting the properties of preference systems. Thus we must have $C = C'$, so by Theorem 26, we also have $A = A'$ and $B = B'$. □

In other words, the preferential content of simple reasons disambiguates the associated range conditional in those cases in which several range conditional expressions are possible. Putting these results together gives us a useful canonical form for simple reasons.

Definition 32 (Canonical forms) *The* canonical form *or representation of a simple reason d is $\emptyset \setminus\!\!\setminus \emptyset \mid\!\vdash \emptyset$ if d is trivial and is the unique interpretation $A \setminus\!\!\setminus B \mid\!\vdash C$ otherwise. We write $d \Rightarrow A \setminus\!\!\setminus B \mid\!\vdash C$ to mean that $A \setminus\!\!\setminus B \mid\!\vdash C$ is the canonical representation of d.*

Clearly, the sets representing the canonical form of a finite simple reason are themselves finite.

The notation $A \setminus\!\!\setminus B \mid\!\vdash C$ permits 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 the representations of previously expressed reasons, even though their meanings change with each enlargement of the set of state components.

The preferences associated with reasons would not deserve much attention if their only purpose was to distinguish one reason from others indicating the same range conditional. In fact, their main purpose lies in choosing among the alternative conclusions demanded by the constitutive meaning of *valid* reasons.

Definition 33 (Validity) *We say that a simple reason $d \in \mathcal{D}$ is* valid *in $S \subseteq \mathcal{D}$ (alternatively, S* validates *d) if the antecedent conditions of its canonical interpretation hold, that is, if $d \Rightarrow A \setminus\!\!\setminus$*

$B \Vdash C$ and $S \in A \setminus\!\!\setminus B$, and that d is invalid *in S (or S* invalidates *d) otherwise. We say that d is* presently valid *in S (or S* presently validates *d) if d is valid in S and $d \in S$, and that d is* presently invalid *in S (or S* presently invalidates *d) if d is invalid in S and $d \in S$. We write $V(d)$ to mean the set of all $S \subseteq \mathcal{D}$ such that d is presently valid in S. Using the same symbol, we write $V(S)$ to mean the set of all $d \in S$ such that d is presently valid in S, and $\overline{V}(S)$ to mean the set $S \setminus V(S)$ of all d presently invalid in S.*

The notions of validity and invalidity employed in (Doyle, 1983c) correspond to the notions of present validity and present invalidity introduced here. These notions permit us to rephrase the preference condition used to identify reasons in Definition 29 as saying that states validating the reason are strictly preferred to states invalidating it when attention is restricted to sets in $[\![d]\!] \cap (A \setminus\!\!\setminus \emptyset)$. The constitutive preference thus provides the simple range conditional with a sense of direction. Note that even presently invalid reasons can be satisfied; that is, we can have d invalid in S even though $S \in [\![d]\!]$. Clearly, if $d \Rightarrow A \setminus\!\!\setminus B \Vdash C$ is valid in $S \in [\![d]\!]$, then $C \subseteq S$. More generally, we use the notion of validity to identify the cases in which the reason *yields* its conclusions.

Definition 34 (Yields) *If $c, d \in \mathcal{D}$ and $S \subseteq \mathcal{D}$, we say that d* (finitely) yields *c* in *S just in case d is a (finite) simple reason valid in S and has canonical representation $A \setminus\!\!\setminus B \Vdash C$ with $c \in C$. We say that S yields c just in case d yields c in S for some $d \in S$. We write $\theta(d, S)$ to mean the set of all elements yielded by d in S, and $\theta(S) \overset{\text{def}}{=} \bigcup_{d \in S} \theta(d, S)$ for the set of all elements yielded by elements of S in S.*

Thus if $d \in V(S)$ and $S \in [\![d]\!]$, we have $\theta(d, S) \subseteq S$. We also note that one never needs conclusions to yield themselves; they are independent of whether the state contains them or not.

Theorem 35 (Conclusion independence) *If $c \in \theta(d, S)$, then $c \in \theta(d, S \setminus \{c\})$.*

PROOF: If $c \in \theta(d, S)$, then d must have canonical representation $A \setminus\!\!\setminus B \Vdash C$ and must be valid in S with $c \in C$. Since $c \notin A$ and $c \notin B$, this means d is also valid in $S \setminus \{c\}$, hence $c \in \theta(d, S \setminus \{c\})$. □

However, we cannot strengthen this to conclude that $c \in \theta(S \setminus \{c\})$ whenever $c \in \theta(S)$, since the only reason yielding c in S may be c itself.

It appears one can alternatively define simple reasons by replacing preference for validity by preference for yielding conclusions, or possibly reduced conclusions. Such an approach would represent an “altruistic” interpretation of reasons as opposed to the “self-interested” approach presented here, with reasons concerned about the “ends” of the reason (the conclusions being held) rather than about the “means” by which these conclusions are held (this reason or some other). However, this alternate definition also appears to introduce unnecessary distinctions among reasons due to antecedents repeated in the conclusions. We do not explore it here.

We may identify several important special forms of simple reasons in terms of the canonical form, defined here and summarized in Table 2.

Definition 36 (Special simple reasons) *If a simple reason has canonical form $A \setminus\!\!\setminus B \Vdash C$, we call the reason* trivial *just in case $C = \emptyset$;* monotonic *just in case $B = \emptyset$;* nonmonotonic *or* defeasible *just in case $B \neq \emptyset$; a* premise *reason just in case $A = B = \emptyset$; and a* basic assumption *just in case $A = \emptyset \neq B$.*

if $A \setminus\!\!\setminus B \Vdash C$ *has*	*the reason is*
$C = \emptyset$	trivial
$B = \emptyset$	monotonic
$A = B = \emptyset$	a premise reason
$B \neq \emptyset$	nonmonotonic or defeasible
$A = \emptyset, B \neq \emptyset$	a basic assumption

Table 2: Special simple reasons

Monotonic reasons yield conclusions that cannot be defeated by enlarging the state, while nonmonotonic reasons yield conclusions that can be defeated by enlarging the state. Trivial reasons never yield any conclusions at all, and are both finite and monotonic. Basic assumption reasons yield defeasible assumptions that do not depend on the presence of any other assumptions in the state, while premise reasons yield conclusions that do not depend on the presence or absence of anything else in the state.

5. Reasoned constitutions

We identified reasons in part in terms of constraints they place on the states in which they appear. But the constraints used to identify reasons capture only part of the full constitutional role of reasons. Recall that the original RMS stability principle requires labeling a node *in* if and only if the recorded reasons contain a valid reason for the node. The meanings imputed to reasons so far ensure that each consequence of a valid reason is in the state, but do nothing to ensure that every element of the state has a valid reason for its presence, much less the RMS groundedness principle that one can identify a set of valid reasons providing a well-founded argument for the element.

The RMS could require grounding arguments for nodes because it distinguished nodes and reasons. In the present setting, we view reasons as state components that may appear as conclusions like other state components. We therefore do not wish to require that *every* state component has a valid reason for its presence, since all reasons themselves would then need reasons themselves, and we would wind up with either infinite states containing infinite regresses or finite states containing circular reasoning structures. To express less comprehensive formulations of this part of the role of reasons, as well as to express the other half of the stability principle, we expand the notions of framings and constitutions employed in the preceding to notions that permit one to consider some elements of states as given and others as derived from the givens.

5.1 Constructive framings

Though quite apt for describing many computational systems, elemental framings consisting simply of the representations explicitly contained in a state are not always the most useful framings of states for analyzing sophisticated reasoning agents, as the set of components explicitly contained in a state mainly serves to represent an "implicitly contained" state description upon which the agent (or more precisely, the agent's action-taking mechanisms) bases its actions. For example, one views some frame-based and taxonomic representational systems as indicating "virtual" representations in addition to those frames or statements actually present in the state (Doyle, 1983b; Fahlman, 1979; Touretzky, 1986); one sometimes also views states consisting of sets of logical statements as indicating their deductive closure (or their closure under a set of inference rules) (Konolige, 1985). To analyze such agents, one does better to frame states in terms of two sets of state components; a set of *manifest* components explicitly contained or represented in the state, and a set of *constructive* components implicitly stored or represented in the state and computed or derived from the manifest components (Doyle, 1989).

Accordingly, we define a *constructive* framing to be a multiattribute framing $\phi : \mathcal{S} \rightarrow \mathcal{S}_{\mathbf{m}} \times \mathcal{S}_{\mathbf{c}}$ interpreting each state as a manifest state in a set $\mathcal{S}_{\mathbf{m}}$ and a constructive state in a set $\mathcal{S}_{\mathbf{c}}$, and write $\phi_{\mathbf{m}} : \mathcal{S} \rightarrow \mathcal{S}_{\mathbf{m}}$ and $\phi_{\mathbf{c}} : \mathcal{S} \rightarrow \mathcal{S}_{\mathbf{c}}$ to mean the corresponding projections of this mapping onto these spaces. The most useful constructive framings for many artificial intelligence systems are *constructive elemental* framings that combine a constructive framing of $\mathcal{S}$ with elemental framings of $\mathcal{S}_{\mathbf{m}}$ and $\mathcal{S}_{\mathbf{c}}$ to yield a framing $\phi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{D}_{\mathbf{m}}) \times \mathcal{P}(\mathcal{D}_{\mathbf{c}})$. We say that a constructive elemental framing is *component compatible* if manifest state components may also be constructive state components, that is, if $\mathcal{D}_{\mathbf{m}} \subseteq \mathcal{D}_{\mathbf{c}}$, and is *extensional* if in addition constructive states always contain their corresponding manifest states, that is, if $\phi_{\mathbf{m}}(s) \subseteq \phi_{\mathbf{c}}(s)$ for every $s \in \mathcal{S}$. We call the framing *uniform* if $\mathcal{D}_{\mathbf{m}} = \mathcal{D}_{\mathbf{c}}$.

Constructive framings may interpret distinct states as containing the same manifest or constructive states. Distinct states sharing their constructive portions need not seem problematic except to lawyers and others seeking to find intentional interpretations of behavior, as we may view the manifest parts of one or both states as containing "redundant" components that do not change the constructive part. But distinct states sharing their manifest parts indicate that the construction process involves nondeterminism, and so may pose problems for mechanizations of agents characterized in this way. Theories of nonmonotonic reasoning provide good examples of this, for one ordinarily permits nonmonotonic reasoning rules to conflict (recall the earlier example involving conflicting rules about Quakers and Republicans), which gives rise to multiple possible interpretations of or constructions from the manifest representations. We say a constructive framing ϕ is *constructively deterministic* just in case $|\phi_{\mathbf{m}}^{-1}(\phi_{\mathbf{m}}(s))| = 1$ for every $s \in \mathcal{S}$, and is *manifestly deterministic* just in case $|\phi_{\mathbf{c}}^{-1}(\phi_{\mathbf{c}}(s))| = 1$ for every $s \in \mathcal{S}$.

5.2 Constructive constitutions

Basic constitutions do not apply directly to state spaces framed with elemental constructive framings, since these map states to pairs of sets rather than to single sets of state components. We therefore extend the notion of basic constitution to that of *constructive* constitution by adding in context-dependent or *contextual* constitutions that describe how manifest and constructive states restrict each other.

Definition 37 (Contextual constitutions) *A* contextual constitution Σ *over* $\mathcal{D}$ *relative to* $\mathcal{D}'$ *is a map* $\Sigma : \mathcal{P}(\mathcal{D}') \rightarrow \Sigma(\mathcal{D})$ *taking each set* $S' \subseteq \mathcal{D}'$ *to a basic constitution* $\Sigma(S')$*. We write the varying parts of a contextual constitution* $\Sigma(S')$ *in the form* $\mathcal{C}(S')$*,* $\vdash(S')$*,* $[\![\,]\!](S')$*, and* $\mathcal{B}(S')$*, and define the set of elements* $\mathcal{A}(\Sigma)$ *to contain just those pairs* $(S', S) \in \mathcal{P}(\mathcal{D}') \times \mathcal{P}(\mathcal{D})$ *such that* $S \in \mathcal{A}(\Sigma(S'))$*.*

Clearly, we may use unstructured contextual constitutions to describe any construction relation whatsoever.

Theorem 38 (General constructions) *If $R \subseteq \mathcal{P}(\mathcal{D}') \times \mathcal{P}(\mathcal{D})$, there is a contextual constitution Σ over $\mathcal{D}$ relative to $\mathcal{D}'$ such that $\mathcal{A}(\Sigma) = R$.*

PROOF: For each $S' \subseteq \mathcal{D}'$, take $\Sigma(S')$ to be the unstructured basic constitution such that $\mathcal{B}(\Sigma(S')) = \{S \subseteq \mathcal{D} \mid (S', S) \in R\}$. Then we clearly have $(S', S) \in \mathcal{A}(\Sigma)$ iff $(S', S) \in R$. □

We may also embed basic constitutions in contextual constitutions over the same set.

Theorem 39 (Embedding constitutions) *If Σ is a basic constitution over $\mathcal{D}$ and Σ' a contextual constitution over $\mathcal{D}$ relative to $\mathcal{D}'$, then there is a contextual constitution Σ'' over $\mathcal{D}$ relative to $\mathcal{D}'$ such that $(S', S) \in \mathcal{A}(\Sigma'')$ iff $(S', S) \in \mathcal{A}(\Sigma')$ and $S \in \mathcal{A}(\Sigma)$.*

PROOF: Define Σ'' to be just like Σ' except that $\mathcal{B}(\Sigma''(S')) = \mathcal{B}(\Sigma'(S')) \cap \mathcal{A}(\Sigma)$ for every $S' \subseteq \mathcal{D}'$. Then we clearly have $(S', S) \in \mathcal{A}(\Sigma'')$ iff $(S', S) \in \mathcal{A}(\Sigma')$ and $S \in \mathcal{A}(\Sigma)$. □

Other embeddings weave the notions of consistency, etc. together in a more natural way, but we do not pursue these here.

We obtain constructive constitutions by combining sub-constitutions describing the manifest and constructive states with contextual sub-constitutions describing the ways each of these relate to each other.

Definition 40 (Constructive constitutions) *A* constructive constitution *Σ over manifest state components $\mathcal{D}_{\mathbf{m}}$ and constructive state components $\mathcal{D}_{\mathbf{c}}$ consists of basic constitutions $\Sigma_{\mathbf{m}}$ over $\mathcal{D}_{\mathbf{m}}$ and $\Sigma_{\mathbf{c}}$ over $\mathcal{D}_{\mathbf{c}}$, together with contextual constitutions $\Sigma_{\mathbf{mc}}$ over $\mathcal{D}_{\mathbf{m}}$ relative to $\mathcal{D}_{\mathbf{c}}$ and $\Sigma_{\mathbf{cm}}$ over $\mathcal{D}_{\mathbf{c}}$ relative to $\mathcal{D}_{\mathbf{m}}$. We define the elements $\mathcal{A}(\Sigma)$ to be those pairs $(S, E) \in \mathcal{P}(\mathcal{D}_{\mathbf{m}}) \times \mathcal{P}(\mathcal{D}_{\mathbf{c}})$ such that $S \in \mathcal{A}(\Sigma_{\mathbf{m}})$, $E \in \mathcal{A}(\Sigma_{\mathbf{c}})$, $(E, S) \in \mathcal{A}(\Sigma_{\mathbf{mc}})$, and $(S, E) \in \mathcal{A}(\Sigma_{\mathbf{cm}})$. If $(S, E) \in \mathcal{A}(\Sigma)$ we also write $S \trianglelefteq E$, and we define the* admissible constructions *of S, written $ACons(S)$, to consist of just those sets E such that $S \trianglelefteq E$.*

Thus the various satisfaction systems of constructive constitutions interpret each manifest (constructive) state component both as a restriction on the manifest (constructive) states in which it can occur, and as a restriction on the ways it can yield (derive from) the constructive (manifest) states.

The four sub-constitutions provided for constructive constitutions clearly represent definitional overkill. Theorem 39 implies the two contextual constitutions may absorb the two basic constitutions, and Theorem 38 implies that each of the contextual constitutions can absorb the resultant restriction of the other in its bounding set. Thus we really only need use one contextual constitution to represent a constructive constitution. In spite of this, however, we keep the definition given above. In the first place, amalgamating all the constitutions in this way may obscure the different restrictions: even though the amalgamation does not change the states so characterized, it may impede exposition and understanding. In the second place, amalgamating the two contextual constitutions into one may yield a constitution with very different locality properties than the original ones, since local logical or satisfaction conditions in the absorbed constitution turn into unstructured general conditions in the absorbing constitution. Again, this does not change the objects so characterized, but it may impede understanding.

These hesitations to amalgamate separate constitutions in the general definition need not, however, prevent us from doing so for the sake of convenience in one important special case, that of the *extensional* constitutions appropriate to extensional framings.

Definition 41 (Extensional constitutions) *An* extensional constitution Σ *is a constructive constitution over manifest state components $\mathcal{D}_{\mathbf{m}}$ and constructive state components $\mathcal{D}_{\mathbf{c}}$ such that $\mathcal{D}_{\mathbf{m}} \subseteq \mathcal{D}_{\mathbf{c}}$ and $S \subseteq E$ whenever $S \trianglelefteq E$, and a* preferential *extensional constitution adds a preference system $\succsim$ over $\mathcal{D}_{\mathbf{c}}$. If $S \trianglelefteq E$, we call E an* admissible extension *of S. We also write $AExts(S)$ to mean $ACons(S)$.*

Thus a preferential extensional constitution yields a preferential basic constitution over the constructive state components for each set of manifest state components. Note that the preference system of a preferential extensional constitution does not depend on the manifest state; a more general treatment would permit contextual variation of it as well.

We can express the extensionality condition of extensional constitutions by means of the contextual constitution $\Sigma_{\mathbf{cm}}$ by taking this constitution to map each context $S \subseteq \mathcal{D}_{\mathbf{m}}$ to the satisfying constitution over $\mathcal{D}_{\mathbf{c}}$ that sets $[\![d]\!](S) = \{E \subseteq \mathcal{D}_{\mathbf{c}} \mid d \in S \rightarrow d \in E\}$ for each $d \in \mathcal{D}_{\mathbf{c}}$. These highly local meanings then imply the global condition that $E \in [\![E]\!](S)$ iff $S \subseteq E$. But rather than go through this exercise when this condition is all that is desired of this contextual constitution, we adopt the shortcut of simply making the bounding set of the other contextual constitution correspond to a subset of the subset relation, that is, making $\mathcal{B}(S) \subseteq \{E \subseteq \mathcal{D}_{\mathbf{c}} \mid S \subseteq E\}$.

We extend the notion of constitutional optimality to extensional constitutions in the obvious way.

Definition 42 (Optimal extensional constitutions) *An extensional constitution Σ over $\mathcal{D}_{\mathbf{m}}$ and $\mathcal{D}_{\mathbf{c}}$ is* optimal *with respect to a preference system $\succsim$ over $\mathcal{D}_{\mathbf{c}}$ just in case E is optimal in $AExts(S)$ with respect to $\succsim$ whenever $S \trianglelefteq E$.*

5.3 Reasoned constitutions

As noted earlier, the meanings associated with reasons in Definition 29 capture half of the RMS stability principle, but do not require that each element of a constructive state possess either a valid reason or a well-founded argument for its presence. We now capture the second half of the basic RMS stability principle, that nodes be labeled in only if they have a valid reason, by applying the requirement of *local groundedness*, or local logical and reasoned derivability, to all constructive state components.

Definition 43 (Local groundedness) *An element $d \in \mathcal{D}$ is* (finitely) locally grounded *in E with respect to $S \subseteq E$ in a preferential basic constitution iff either*

1. *$d \in S$,*

2. *$E' \setminus \{d\} \vdash d$ for some (finite) $E' \subseteq E$, or*

3. *$E \setminus \{d\}$ (finitely) yields d.*

When this happens, we write $S <_{\mathbf{l}}^{d} E$. We write $S <_{\mathbf{l}} E$ just in case $S <_{\mathbf{l}}^{d} E$ for each $d \in E$.

If we say d is *given* in E with respect to $S \subseteq E$ just in case $d \in S$, then a state component is locally grounded just in case it is either given, logically derivable from other state components, or yielded by some other state component. We then define reasoned constitutions as follows.

Definition 44 (Reasoned constitutions) *A* reasoned constitution Σ *is a uniform preferential extensional constitution such that every element of every admissible extension is locally grounded with respect to the manifest state, that is, $S <_{\mathbf{l}} E$ with respect to $\Sigma(S)$ and the preference system whenever $S \trianglelefteq E$.*

We can express this condition of universal local grounding in a variety of ways. The most natural and local formulation places it in the contextual constitution $\Sigma_{\mathbf{cm}}$ to require that for every $d \in \mathcal{D}_{\mathbf{c}}$ we have

$$[\![d]\!]_{\mathbf{cm}} \subseteq \{(S, E) \in \mathcal{P}(\mathcal{D}_{\mathbf{m}}) \times \mathcal{P}(\mathcal{D}_{\mathbf{c}}) \mid d \in E \rightarrow S <_{\mathbf{l}}^{d} E\}.$$

That is, each element requires that it be locally grounded whenever it appears in an admissible extension.

The definition of reasoned constitution stipulates only that constructive state components must be locally grounded, not that they must possess any more complex arguments for their presence. Since constructive constitutions already require states to satisfy the reasons they contain, this definition identifies reasoned constitutions as those which enforce the RMS stability principle. We omit any requirement corresponding to the RMS grounding principle from the definition of reasoned constitutions so that we may consider reasoning systems that employ a variety of patterns of grounding of conclusions. For example, because humans do not appear to remember or enforce reasons for most of their beliefs, some authors (e.g., (Gärdenfors, 1990; Harman, 1986); cf. (Doyle, 1992)) criticize the RMS for requiring extensive grounding of all elements (although people do appear excellent at effortlessly constructing rationalizations or supporting arguments—sometimes spurious—for their beliefs (Gazzaniga, 1985)). Our definition of reasoned constitution accommodates these concerns both by taking manifest state components as givens needing no supporting reasons (though not forbidding such supporting reasons), and by encompassing reasoned constitutions that require different levels of grounding for different state components, for example, constitutions requiring only minimal local grounding of elements of long-term memory, for which people seem especially prone to ignore reasons, but requiring stricter grounding patterns for elements of short-term memories, for which people seem more aware of their reasons for thinking things. We thus identify reasoned constitutions by the minimal requirement that everything not given must have a reason or immediate derivation, and then examine special reasoned constitutions that impose stronger grounding requirements for some or all constructive state components.

The most natural stronger grounding condition, which we simply call *groundedness* (or *strict* groundedness) strengthens the immediate derivability of local groundedness to derivability by an argument mixing entailment and reasons that traces back to givens, such that each step of the argument makes the same assumptions as all the others.

Definition 45 (Groundedness) *We say that e is* (finitely) grounded *in E with respect to $S \subseteq E$ in a preferential basic constitution iff 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. *$d \in S$,*
2. *$G_{<d} \vdash d$, where $G_{<d} \stackrel{\text{def}}{=} \{g \in G \mid g <_G d\}$, or*
3. *there is some (finite) simple reason $f \in G_{<d}$ yielding d in E such that $f \Rightarrow A \setminus\!\!\setminus B \Vdash C$ and $A \subseteq G_{<f}$.*

When this happens, we write $S <_{\mathbf{g}}^{d} E$ ($S <_{\mathbf{f}}^{d} E$). We write $S <_{\mathbf{g}} E$ ($S <_{\mathbf{f}} E$) just in case $S <_{\mathbf{g}}^{d} E$ ($S <_{\mathbf{f}}^{d} E$) for each $d \in E$.

Note well that the qualifications of reasons yielding conclusions must be checked against the whole set E, while the antecedents must precede the reason in the grounding order. Also, as one might expect, groundedness implies local groundedness.

Theorem 46 (Local grounding) *If e is (finitely) grounded in E with respect to $S \subseteq E$, then e is (finitely) locally grounded in E with respect to S.*

PROOF: Suppose e is (finitely) grounded in E with respect to $S \subseteq E$ with (finite) grounding set $G \subseteq E \setminus \{e\}$. Then by definition, either $e \in S$, or there is a (finite) set $A \subseteq G$ such that $A \vdash \{e\}$, or some presently valid (finite) reason in $E \setminus \{e\}$ yields e. □

The notions of groundedness and local groundedness, however, represent extremes along a dimension rather than the only two possibilities. For example, as (Doyle, 1983b; Doyle & Wellman, 1990) suggest one might employ intermediate notions of groundedness with respect to grounding neighborhoods or locales defined to be strict grounding for elements of the neighborhoods or locales, but only local grounding for elements outside them. This paper does not pursue intermediate notions of groundedness or nonuniform degrees of grounding of admissible extensions. We instead focus on the special cases of uniformly grounded reasoned constitutions, named as follows.

Definition 47 (Special reasoned constitutions) *We say a reasoned constitution is* (finitely) grounded *just in case it requires every element of every admissible extension to be (finitely) grounded with respect to the manifest state.*

6. Grounded extensions

We now turn to analyzing the structure of admissible extensions in grounded reasoned constitutions, and for simplicity of analysis ignore the structure of manifest states. The manifest states of RMS and many other systems really consist of sets of stipulated state components, which may be stipulated one by one, resulting in arbitrary sets of stipulated components. The reasoned constitutions corresponding to these systems thus employ trivial constitutions for manifest states and purely extensional contextual constitutions for manifest states relative to constructive states, so the analysis of their admissible extensions mirrors the analysis of *extensions* that ignore the internal structure of manifest states to focus only on the constitutions of constructive states and the contextual constitutions specifying how each constructive element is grounded in the manifest state.

Definition 48 (Extensions) *If Σ is a basic constitution over $\mathcal{D}$, we say that E is an* extension *of (or a* state extending*) $S \subseteq \mathcal{D}$ iff $S \subseteq E \in \mathcal{A}(\Sigma)$, and write $S \lhd E$ to indicate that E extends S, and $Exts(S)$ to denote the set of extensions of S. If Σ is an extensional constitution over $\mathcal{D}_{\mathbf{m}}$ and $\mathcal{D}_{\mathbf{c}}$, we say that E is an extension of S iff $S \in \mathcal{A}(\Sigma_{\mathbf{m}})$, $E \in \mathcal{A}(\Sigma_{\mathbf{c}})$, and $S \subseteq E$. If $S \lhd E$ and $S <_{\mathbf{g}} E$ ($S <_{\mathbf{l}} E$, $S <_{\mathbf{f}} E$), we write $S \lhd_{\mathbf{g}} E$ ($S \lhd_{\mathbf{l}} E$, $S \lhd_{\mathbf{f}} E$). We write $GExts(S)$, $LExts(S)$, and $FGExts(S)$ to mean, respectively, the sets of grounded, locally grounded, and finitely grounded extensions of S.*

Finitely grounded extensions are grounded extensions by definition, and we see that grounded extensions are locally grounded as an immediate corollary of Theorem 46.

6.1 Stratification

We first examine an important alternate characterization of the notions of grounded and finitely grounded extensions. Here and in the following we assume a fixed reasoned constitution.

Definition 49 (Levels) *Let $S, E \subseteq \mathcal{D}$. Then the sequence $\langle \Lambda_\alpha \rangle$ (α an ordinal), the levels from S in E, are defined for all ordinals by*

1. $\Lambda_0(S, E) = S$,
2. $\Lambda_{\alpha+1}(S, E) = \Theta(\Lambda_\alpha(S, E)) \cup \theta(\Lambda_\alpha(S, E))$, *and*
3. $\Lambda_\lambda(S, E) = \bigcup_{\alpha<\lambda} \Lambda_\alpha(S, E)$ *for limit ordinals* λ.

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

Note how each level includes the deductive closures of the preceding levels of inference via valid reasons. We also easily observe that if $\alpha \leq \beta$, then $S \subseteq \Lambda_\alpha(S, E) \subseteq \Lambda_\beta(S, E) \subseteq \Lambda(S, E)$, and that if $\Lambda_\alpha = \Lambda_{\alpha+1}$, then $\Lambda_\alpha = \Lambda$ (here and in the following, we sometimes omit the level parameters (S, E) when the meaning is clear).

Theorem 50 *If $\alpha > |\mathcal{D}|$, then $\Lambda(S, E) = \Lambda_\alpha(S, E)$.*

PROOF: If $\mathcal{D}$ 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 $\Lambda = \Lambda_\alpha$. □

As this result suggests, the division of extensions into levels naturally leads to ranking elements according to the first level at which they appear.

Definition 51 (Rank) *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.*

With the notion of rank, we first observe that the sum of all levels is closed.

Lemma 52 (Closure) *For every $S, E \subseteq \mathcal{D}$, $\Lambda(S, E) = \Theta(\Lambda(S, E))$.*

PROOF: Suppose $A \subseteq \Lambda$. A has rank, say α, so if $A \vdash B$, then $B \subseteq \Lambda_{\alpha+1} \subseteq \Lambda$. Since $\Lambda \subseteq \Theta(\Lambda)$, we have $\Lambda = \Theta(\Lambda)$. □

In addition, extensions contain the sum of all their levels.

Lemma 53 (Containment) *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 $e \in \Theta(\Lambda_\beta)$, then $e \in E$ since E is closed. If $e \notin \Theta(\Lambda_\beta)$ there is a $d \in \Lambda_\beta$ with $d \Rightarrow A \setminus\!\!\setminus B \mathrel{\|\!\!-} C$, $A \subseteq \Lambda_\beta$, $E \subseteq \overline{B}$, and $e \in C$. Since E is satisfying, this means $C \subseteq E$, so $e \in E$. Hence $\Lambda_\alpha \subseteq E$, so $\Lambda \subseteq E$. □

Moreover, grounded extensions equal the sum of all their levels.

Theorem 54 (Stratification) *If $S \lhd_{\mathbf{g}} E$ then $\Lambda(S, E) = E$.*

PROOF: Suppose $S \lhd_{\mathbf{g}} E$. Since $S \lhd E$, by Lemma 53 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 closed. Otherwise there is a $d \in G$ such that $d \Rightarrow A \setminus\!\!\setminus B \mathrel{\Vdash} C$, $A <_G d <_G f$, $E \subseteq \overline{B}$, 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 $E = \Lambda$. □

Obviously, If $S \lhd_{\mathbf{g}} E$ and $\alpha > |E|$, then $E = \Lambda_\alpha(S, E)$, and if $S \lhd_{\mathbf{g}} E$ and α is the rank of E, then $E = \Lambda_\alpha(S, E)$. We thus easily see that finitely grounded extensions equal the sum of their finite levels.

Corollary 55 (Countable stratification) *If $S \lhd_{\mathbf{f}} E$, then $E = \Lambda_\omega(S, E)$.*

PROOF: Let $S \lhd_{\mathbf{f}} E$ 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 53, $E = \Lambda_\omega(S, E)$. □

The second main result shows that fixed points of the level operator are grounded extensions, assuming they are consistent, bounded, and satisfy all their non-simple-reason elements (by this we mean that the set satisfies each of its elements that is not a simple reason).

Theorem 56 (Fixed point) *If E is consistent, bounded, satisfies its non-simple-reason elements and $E = \Lambda(S, E)$, then $S \lhd_{\mathbf{g}} E$.*

PROOF: Suppose E is consistent, bounded, satisfies its non-simple-reason elements, and $\Lambda = E$. Since $S \subseteq \Lambda$, $S \subseteq E$. Let $e \in E$ with $e \Rightarrow A \setminus\!\!\setminus B \mathrel{\Vdash} C$, 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 $E \subseteq \overline{B}$ as well, then $C \subseteq \Lambda_{\alpha+1} \subseteq E$. Thus E is satisfying. Similarly, if $A \vdash B$, $A \subseteq \Lambda$, and A has rank α, then $B \subseteq \Lambda_{\alpha+1} \subseteq \Lambda$. Thus E is closed, and therefore an extension. 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 the rank of a does not exceed the rank of 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 $e \in \Theta(\Lambda_\beta)$, then there is some $G \subseteq \Lambda_\beta$ such that $G \vdash \{e\}$, and otherwise there is some $d \in \Lambda_\beta$ such that $d \Rightarrow A \setminus\!\!\setminus B \mathrel{\Vdash} C$, $A \subseteq \Lambda_\beta$, $E \subseteq \overline{B}$, 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 $S \lhd_{\mathbf{g}} E$. □

Putting these results together, we arrive at our main characterization of grounded extensions.

Corollary 57 (Grounded extensions) *If E is consistent, bounded, and satisfies its non-simple-reason elements, then $S \lhd_{\mathbf{g}} E$ iff $E = \Lambda(S, E)$.*

From this we derive a corresponding characterization of finitely grounded extensions.

Theorem 58 (Finitely grounded extensions) *If E is consistent, bounded, satisfies its non-simple-reason elements, $\vdash$ is compact, and every simple reason in $\mathcal{D}$ is finite, then $S \lhd_{\mathbf{f}} E$ iff $E = \Lambda_\omega(S, E)$.*

PROOF: Suppose E is consistent, bounded, satisfies its non-simple-reason elements, $\vdash$ is compact, and every simple reason in $\mathcal{D}$ is finite. By Corollary 55, we need only show that $\Lambda_\omega = E$ implies $S \lhd_{\mathbf{f}} E$. 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 $e \in \mathcal{D}$, $e \in \Lambda_{\alpha+1} \setminus \Lambda_\alpha$. Since α is minimal, $\Lambda_\omega = \Lambda_\alpha$, for otherwise $\Lambda_\omega = \Lambda_{\omega+1}$ and hence $\Lambda_\omega = \Lambda$. If $e \in \Theta(\Lambda_\omega)$ 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 $e \notin \Theta(\Lambda_\omega)$, then by construction, there is some $f \in \Lambda_\omega$, $f \Rightarrow A \mathbin{\backslash\!\backslash} B \Vdash C$, $A \subseteq \Lambda_\omega$, $E \subseteq \overline{B}$, 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 56 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 $e \in \Theta(\Lambda_\alpha)$, then there is a finite $G \subseteq \Lambda_\alpha$ such that $G \vdash \{e\}$. If $e \notin \Theta(\Lambda_\alpha)$, then by construction there is some $f \in \Lambda_\alpha$ with $f \Rightarrow A \mathbin{\backslash\!\backslash} B \Vdash C$, $A \subseteq \Lambda_\alpha$, $E \subseteq \overline{B}$, 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. □

Theorem 59 (Monotonic determinism) *If $\mathcal{D}$ contains only monotonic simple reasons, every subset of $\mathcal{D}$ has a unique grounded extension.*

PROOF: Suppose each element of $\mathcal{D}$ is a monotonic simple reason, and let $S \subseteq \mathcal{D}$. Consider $E = \Lambda(S, \emptyset)$. Since all simple 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 $S \lhd_{\mathbf{g}} E$ by Theorem 56. Now if $S \lhd_{\mathbf{g}} E'$, then $E' = \Lambda(S, E') = E$ by the previous observation, so E is the only grounded extension. □

As a corollary, if $\mathcal{D}$ contains only finite monotonic simple reasons, every subset of $\mathcal{D}$ has a unique finitely grounded extension.

We also find that grounded extensions are minimal among locally grounded extensions.

Theorem 60 (Grounded minimality) *If $S \lhd_{\mathbf{g}} E$, $S \lhd_{\mathbf{l}} E'$, and $E' \subseteq E$, then $E = E'$.*

PROOF: Suppose $S \lhd_{\mathbf{g}} E$, $S \lhd_{\mathbf{l}} E'$, 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 $e \in \Theta(\Lambda_\beta(S, E))$, then $e \in E'$, and otherwise there is a $d \in \Lambda_\beta(S, E)$ with $d \Rightarrow A \mathbin{\backslash\!\backslash} B \Vdash C$, $A \subseteq \Lambda_\beta(S, E)$, $E \subseteq \overline{B}$, and $e \in C$. But since $E' \subseteq E \subseteq \overline{B}$, this means $e \in \Lambda_\alpha(S, E')$. Hence $\Lambda(S, E) \subseteq \Lambda(S, E')$. But by Theorem 54 and Lemma 53, $E = \Lambda(S, E) \subseteq \Lambda(S, E') \subseteq E'$, hence $E = E'$. □

Obviously, this result means that grounded extensions are minimal among grounded extensions, and that finitely grounded extensions are minimal among finitely grounded extensions.

Our earlier treatment (Doyle, 1983c) also explores a different means for decomposing extensions. Where stratification partitions elements according to how many steps (logical or reasoned) they take to derive, one may instead partition the underlying set $\mathcal{D}$ according to which state components each state component *mentions*, in the following sense.

Definition 61 (Reasoned decomposition) *Suppose $S \subseteq \mathcal{D}$ and $d \in \mathcal{D}$. The set S* mentions *d iff for some $e \in S$, $e \Rightarrow A \setminus\!\!\setminus B \Vdash C$ 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 $S \setminus A$ have disjoint mention sets. A set $S \subseteq \mathcal{D}$ is called* simple *iff S has no isolated subsets other than itself and $\emptyset$. 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 $d \Rightarrow A \setminus\!\!\setminus B \Vdash C$, then $A, B, C \subseteq \mathcal{U}(S)$.*

For example, S and $\emptyset$ are isolated subsets of S, and if $\mathcal{D} = \{a\}$, then $\mathcal{D}$ is simple, while if $\mathcal{D} = \{a, b\}$ where both elements have trivial interpretations, then $\mathcal{D}$ is not simple, since each of $\{a\}$ and $\{b\}$ is. In addition, if $\mathcal{D}$ is finite, then every subset has a finite universe. Moreover, if $S \subseteq \mathcal{D}$ has a finite universe, then all simple reasons in S are finite. Sets with disjoint universes have disjoint mention sets, though the converse need not be true.

These definitions permit an interesting analysis of the structure, computability, and complexity of extensions when the information system under consideration is trivial, and in particular allow one to avoid assumptions about the finiteness of $\mathcal{D}$ in favor of assumptions of finite universes of sets of manifest components. We do not know if these definitions may be generalized to provide useful results when the constitution involves a nontrivial information system.

6.2 Optimality

We saw in Section 4.1 that simple range conditionals, and therefore simple reasons, prove interesting with respect to the defined notion of strong present satisfaction preference. However, we defined simple reasons with respect to that portion of an arbitrary preference system related to the validity of reasons. Investigating optimality with respect to this portion of the given preference system provides further insight into conclusions drawn from simple reasons.

Definition 62 (Validity preferences) *The* pure validity *preference system corresponding to a basic constitution Σ over $\mathcal{D}$ is the preference system $(\mathcal{D}, \precsim)$ such that for each $d \in \mathcal{D}$ and $S, S' \subseteq \mathcal{D}$, we have $S \sim_d S'$ iff either $S = S'$ or d is not a simple reason, and have $S \succ_d S'$ just in case d is a simple reason valid in S and invalid in S'. The* weak present validity *preference system instead defines $S \succ_d S'$ just in case d is a simple reason presently valid in S but not presently valid in S' (either $d \notin S'$ or d is presently invalid in S'). The* strong present validity *preference system instead defines $S \succ_d S'$ just in case d is a simple reason presently valid in S and presently invalid in S'. We call any set optimal in some range with respect to the pure, weak, or strong present validity preference systems (respectively)* purely, weakly, or strongly present validity optimal *in the range.*

Thus a set S is weakly present validity optimal in R iff for each $d \in \mathcal{D}$ and $S' \in R$, if S' presently validates d but S does not, then there is some e presently valid in S but not presently valid in S'; phrased differently, if $d \in V(S')$ but $d \notin V(S)$, then there is some $e \in V(S)$ with $e \notin V(S')$.

Similarly, a set S is strongly present validity optimal in R iff for each $d \in \mathcal{D}$ and $S' \in R$, if S' presently validates d while S presently invalidates d, then there is some e presently valid in S but presently invalid in S', or equivalently, if $\overline{V}(S') \cap V(S) \neq \emptyset$ whenever $V(S') \cap \overline{V}(S) \neq \emptyset$. Obviously, if S is strongly present validity optimal in R, then S is weakly present validity optimal in R. The notions of weakly present validity optimality and strongly present validity optimality correspond to the notions of validity optimality and strong validity optimality of (Doyle, 1983c).

As one might expect, since the varieties of validity optimality involve an arbitrary preference system which may bear no relation to satisfaction systems, validity optimality proves very different from satisfaction optimality. For example, satisfying sets need not be weakly present validity optimal among satisfying sets. To see this, let $\mathcal{D} = \{d\}$, and $[\![d]\!] = \mathcal{P}(\mathcal{D})$. Then both $\emptyset$ and $\mathcal{D}$ are satisfying, $\mathcal{D}$ presently validates d, but $\emptyset$ presently validates no element not in $\mathcal{D}$, so $\emptyset$ is not weakly present validity optimal in $\mathcal{P}(\mathcal{D})$. In addition, unsatisfying sets may be weakly present validity optimal in $\mathcal{P}(\mathcal{D})$. We see this by letting $\mathcal{D} = \{d, e\}$, $d \Rightarrow \emptyset \mathbin{\backslash\!\backslash} \{d\} \mathrel{\|\!\!-} \{e\}$, and $e \Rightarrow \emptyset \mathbin{\backslash\!\backslash} \{e\} \mathrel{\|\!\!-} \{d\}$. Then $\{d\}$ and $\{e\}$ are unsatisfying, but every set in $\mathcal{P}(\mathcal{D})$ is validity optimal since no set validates any of its own elements.

We now use these optimality notions to show that grounded extensions make as many assumptions as possible given the reasons they contain.

Theorem 63 (Weak optimality) *Grounded constitutions are optimal with respect to weak present validity preferences.*

PROOF: Suppose, by way of contradiction, that $S \trianglelefteq E$, $S \trianglelefteq E'$ and that there is a d is valid in E' but not valid in E, but 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 $V(E) \subseteq V(E')$. The element d shows $E \neq E'$, so by the minimality of E and E' among grounded extensions, $E \setminus E' \neq \emptyset$. Let $e \in E \setminus E'$. Since $S \subseteq E'$, $e \notin S$. But since E is locally grounded, 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 weakly present validity optimal in $AExts(S)$, which means that the constitution is optimal with respect to weak present validity preference. □

We interpret this result as indicating the preferential orthogonality of grounded extensions, an orthogonality that appears in nonmonotonic and default logics as the logical inconsistency of alternative extensions. We take this to demonstrate that one can sensibly speak of psychological incompatibility without requiring notions of logical inconsistency. We do not know whether preferential orthogonality characterizes grounded extensions, that is, whether any set is validity optimal among admissible extensions is itself an admissible extension. We can, however, strengthen the result somewhat for finitely grounded constitutions.

Theorem 64 (Strong optimality) *Finitely grounded constitutions in which all reasons are finite are optimal with respect to strong present validity preferences.*

PROOF: Suppose $S \trianglelefteq E$ in a finitely grounded constitution in which all reasons are finite, and that E is not strongly present validity optimal. Then there is some E' such that $S \trianglelefteq E'$ and some $d \in E$ 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' among grounded extensions, $E \setminus E' \neq \emptyset$. 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 $e \in \Lambda_{\alpha+1}(S, E) \setminus \Lambda_{\alpha+1}(S, E')$. Then there is some $f \in \Lambda_\alpha$ with $f \Rightarrow A \mathbin{\backslash\backslash} B \Vdash C$, $A \subseteq \Lambda_\alpha$, $E \subseteq \overline{B}$, 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 present validity optimal in $AExts(S)$. □

7. Conclusion

This paper presented elements of a mathematical theory of nonmonotonic reasoning based on concepts chosen to isolate the fundamental structures of the subject rather than to exhibit a conventional logical appearance. The approach taken here makes no special assumptions about the makeup of minds, but instead interprets given structures naturalistically, identifying reasons and their conclusions by the roles they play in the makeup of psychological states. It avoids special assumptions about how minds realize or compute these roles, so encompassing the computationally trivial activities of reason maintenance systems as well as the unbounded reasoning powers of ideal rational agents, and treats nonmonotonic reasoning as applicable to all sorts of mental attitudes, not as forms of reasoning specific to or justified only for beliefs, thus sidestepping the artificial consistency requirements attendant in logical encodings that limit reasoning about conflicting information or attitudes. The resulting theory thus exhibits the essential properties of traditional logical theories of nonmonotonic reasoning, but makes optional the more specific (or dubious) properties of the logical theories.

By avoiding the unnecessary concomitants of logical encodings of reasoning, the approach here applies directly to a wide variety of psychological structures. While the particular formulation presented here may not adequately serve the analysis of all interesting psychological structures, I believe it provides a good basis for studying an important range of questions, and that many further investigations will require only additions and refinements to the concepts presented here rather than wholesale replacement.

The formulation of ideas presented here arose through the deliberate pursuit of a mathematical understanding of the subject, and through deliberate practice of the essentially mathematical methodology of rational psychology sketched in the introduction and at somewhat greater length in (Doyle, 1983d). Rational psychology offers benefits to all of artificial intelligence and the cognitive sciences, not just to the theory of nonmonotonic reasoning. The preceding development exhibits some of the benefits of this approach, and I hope the reader will have come to agree on the merits of this project, even if unsatisfied with the specific formulations presented above.

Peregrinus expectavi pedes meos in cymbalis.
S. Prokofiev, *Alexander Nevsky*

Acknowledgements

Preparation of this paper was supported in part by the USAF Rome Laboratory and ARPA under contract F30602-91-C-0018, and by National Institutes of Health Grant No. R01 LM04493 from the National Library of Medicine. The author gratefully acknowledges this support, and has also benefited over the years from the invaluable advice and encouragement of Joseph Schatz, Gerald Sussman, Dana Scott, and Richmond Thomason, as well as the productive environments of MIT, Stanford, and Carnegie Mellon University.

References

Abelson, H. (1978). Towards a theory of local and global in computation. *Theoretical Computer Science*, *6*, 41–67.

de Kleer, J., Doyle, J., Steele, G. L. Jr., & Sussman, G. J. (1977). AMORD: Explicit control of reasoning. In *Proceedings of the ACM Symposium on Artificial Intelligence and Programming Languages*, pp. 116–125.

Doyle, J. (1976). The use of dependency relationships in the control of reasoning. Tech. rep. Working Paper 133, MIT AI Laboratory.

Doyle, J. (1979). A truth maintenance system. *Artificial Intelligence*, *12*(2), 231–272.

Doyle, J. (1980). A model for deliberation, action, and introspection. Ai-tr 581, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA, 02139.

Doyle, J. (1983a). Admissible state semantics for representational systems. *IEEE Computer*, *16*(10), 119–123.

Doyle, J. (1983b). A society of mind: Multiple perspectives, reasoned assumptions, and virtual copies. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp. 309–314.

Doyle, J. (1983c). Some theories of reasoned assumptions: An essay in rational psychology. Tech. rep. 83-125, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Doyle, J. (1983d). What is rational psychology? toward a modern mental philosophy. *AI Magazine*, *4*(3), 50–53.

Doyle, J. (1985). Circumscription and implicit definability. *Journal of Automated Reasoning*, *1*, 391–405.

Doyle, J. (1988). Artificial intelligence and rational self-government. Tech. rep. CS-88-124, Carnegie-Mellon University Computer Science Department.

Doyle, J. (1989). Constructive belief and rational representation. *Computational Intelligence*, *5*(1), 1–11.

Doyle, J. (1992). Reason maintenance and belief revision: Foundations vs. coherence theories. In Gärdenfors, P. (Ed.), *Belief Revision*, pp. 29–51. Cambridge University Press, Cambridge.

Doyle, J., & Wellman, M. P. (1990). Rational distributed reason maintenance for planning and replanning of large-scale activities. In Sycara, K. (Ed.), *Proceedings of the DARPA Workshop on Planning and Scheduling*, pp. 28–36 San Mateo, CA. Morgan Kaufmann.

Doyle, J., & Wellman, M. P. (1991). Impediments to universal preference-based default theories. *Artificial Intelligence*, *49*(1-3), 97–128.

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

Gärdenfors, P. (1990). The dynamics of belief systems: Foundations vs. coherence theories. *Revue Internationale de Philosophie*, *172*, 24–46.

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

Ginsberg, M. L. (1991). The computational value of nonmonotonic reasoning. In Allen, J., Fikes, R., & Sandewall, E. (Eds.), *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pp. 262–268 San Mateo, CA. Morgan Kaufmann.

Harman, G. (1986). *Change in View: Principles of Reasoning*. MIT Press, Cambridge, MA.

James, W. (1892). *Psychology*. Henry Holt & Co., New York.

Konolige, K. (1985). Belief and incompleteness. In Hobbs, J. R., & Moore, R. C. (Eds.), *Formal Theories of the Common-Sense World*, pp. 359–403. Ablex, Norwood.

Makinson, D., & Gärdenfors, P. (1991). Relations between the logic of theory change and the nonmonotonic logic. In Fuhrmann, A., & Morreau, M. (Eds.), *The Logic of Theory Change*, pp. 185–205. Springer-Verlag, Berlin.

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B., & Michie, D. (Eds.), *Machine Intelligence 4*, pp. 463–502. Edinburgh University Press.

McCarthy, J. (1977). Epistemological problems of artificial intelligence. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pp. 1038–1044.

McCarthy, J. (1980). Circumscription — a form of non-monotonic reasoning. *Artificial Intelligence*, *13*(1), 27–38.

McDermott, D. (1978). Planning and acting. *Cognitive Science*, *2*, 71–109.

McDermott, D., & Doyle, J. (1980). Non-monotonic logic—I. *Artificial Intelligence*, *13*, 41–72.

Miller, G. A. (1986). Dismembering cognition. In Hulse, S. H., & Green, Jr., B. F. (Eds.), *One Hundred Years of Psychological Research in America*, pp. 277–298. Johns Hopkins University Press, Baltimore.

Minsky, M. (1965). Matter, mind, and models. In *Proceedings of the IFIP Congress*, pp. 45–49.

Minsky, M. (1980). K-lines: a theory of memory. *Cognitive Science*, *4*, 117–133.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, *13*, 81–132.

Sandewall, E. (1972). An approach to the frame problem, and its implementation. In *Machine Intelligence 7*, pp. 195–204. University of Edinburgh Press.

Scott, D. S. (1982). Domains for denotational semantics. In Nielsen, M., & Schmidt, E. M. (Eds.), *Automata, Languages, and Programming: Ninth Colloquium*, Vol. 140 of *Lecture Notes in Computer Science*, pp. 577–613 Berlin. Springer-Verlag.

Sussman, G. J., Winograd, T., & Charniak, E. (1971). Micro-planner reference manual, (revised). Aim 203A, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA, 02139.

Touretzky, D. S. (1986). *The Mathematics of Inheritance Systems*. Morgan Kaufman, Los Altos, CA.

Truesdell, C. (1958). Recent advances in rational mechanics. *Science*, *127*, 729–739.

Truesdell, C. (1977). *A First Course in Rational Continuum Mechanics*, Vol. 1. Academic Press, New York.