Reset reproduction of CMU Computer Science report CMU-CS-83-127. Published in *IJCAI'83*, pp. 309–314. Reprinted July 1994. Current address: MIT Laboratory for Computer Science, Cambridge, Massachusetts.

A Society of Mind

Multiple perspectives, reasoned assumptions, and virtual copies

Jon Doyle

Computer Science Department
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213
U.S.A.

Abstract: Little is clearly understood about the similarities, differences, and comparative computational and representational advantages of the many proposals extant for organizing minds into collections of "mental subagents." Using a new mathematical framework for exactly specifying the structure of mental organizations, we formulate separately the ideas of multiple perspectives, reasoned assumptions, and virtual copies. When combined, these notions form a common backbone for systems as diverse as CONLAN, NETL, and FOL, and show many particular characteristics of those systems results of the "language of thought" adopted for representing the contents of mental subagents. The framework also suggests connections between the "strengths" of mental attitudes, the ambiguity of "self," and the possibilities for self-omniscience.

This paper will be presented at IJCAI-83.

I thank Joseph Schatz for advice, and Ben Cohen for reading a distant ancestor of this paper. This research was supported by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionics Laboratory under Contract F33615-81-K-1539. The views and conclusions contained in this document are those of the author, and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Government of the United States of America.

§**1.** Several recent proposals in artificial intelligence reformulate ancient doubts about the reality of the "self" by explaining or constructing agents in terms of a collection of interacting, simpler subagents. Some of these proposals discuss the agent's actions without reference to any "self" at all, and others presume "selfhood" to flit epiphenomenally from subagent to subagent as dictated by needs to communicate with the external world or to assign credit or blame for actions. While thinkers throughout history have occasionally doubted on philosophical grounds the common "single-agent" view of the human mind, the new proposals suggest that there may be computational difficulties inherent in single-agent psychologies that are only overcome by the multi-agent viewpoint. Minsky has called this approach the "society of mind." In some proposals the subagents comprising the society are numerous, very simple, neurologically conceived mechanisms. Other proposals suggest more complex subagents, ranging from the coroutine collections of "heterarchy," to the "knowledge sources" of production systems, to frame systems, to collections of mutually referring logical theories, to Freud's committee of id, ego, and superego, and modern split-brain theories, in which the complexities of the subagents rival that of the agent as a whole. We need not view these proposals as mutually exclusive, if we can subdivide subagents into sub-subagents, but questions like this are difficult to pursue without reasonably precise characterizations of the different sorts of subagents to be related. More specifically, the knowledge representation literature is filled with proposals for complex organizations based on widely differing sorts of "languages of thought," such as logic (FOL [Weyhrauch 1980]), list structures and rational algebraic functions (CONLAN [Sussman and Steele 1980]), nodes and links (NETL [Fahlman 1979]), etc. Although their abstract structures seem related, there is little hope for understanding the relations among these proposals and for making rapid further progress without clearly formulating the underlying ideas separately and then analyzing their range of combinations. Toward this end, we present a mathematical framework for exactly specifying the structure of mental societies. Since the framework is fairly general, we illustrate it by characterizing a particular society of mind which incorporates three often-proposed capabilities of subagents and relations between subagents, namely multiple perspectives, reasoned assumptions, and virtual copies. While these characteristics of societies are sometimes thought to require the use of logical or quasi-logical languages as systems of representation, our formulation makes few structural or representational demands, and so permits use of any desired system of representation (including logical languages) in which the few required structures may be encoded. For example, we can reconstruct FOL and CONLAN at the end of our formulation largely by choosing logical or LISP-like languages for the contents of mental subagents.

The mathematical framework is developed and otherwise applied in [Doyle 1982] and [Doyle 1983b]. While I formulate the particular society here to generalize the organization suggested in my thesis ([Doyle 1980]), the ideas involved have an older, wider history, and I have worked to incorporate the insights of Johan de Kleer, Merrick Furst, Kurt Konolige, Marvin Minsky, Brian Smith, Richard Stallman, Guy Steele, Gerald Sussman, David Touretzky, and Richard Weyhrauch into this exposition.

§**2.** Researchers frequently motivate proposed decompositions of mind with concerns about self-knowledge, that is, information and mechanisms the agent employs to understand, predict, control, and modify its structures and actions. Although specific tasks appear amenable to specific solutions, students of the broad problems of representation, decision-making, and learning come to appreciate the utility, if not importance, of self-knowledge in adaptive agents. Artificial intelligence studies many sorts of self-knowledge, but for brevity we consider only three.

One commonly studied sort of self-knowledge involves multiple coreferential representations. Since artificial intelligence proposals often suppose representational agents, individual representations and their relations form natural objects of self-knowledge. Since the feasible mechanization of thinking demands concern for the difficulty of reaching conclusions and solving problems, one of the most studied relations between representations is the ease of thinking about something in terms of one representation relative to the ease of thinking about it in terms of an alternative representation. Minsky emphasizes reformulation or representation switching as the heart of problem solving; Bobrow and Winograd make similar oppor-

tunism the basis of KRL; and SUSSMAN and STEELE illustrate the inferential importance of interactions between multiple coreferential representations. Thus useful sorts of self-knowledge include the possible alternatives to a particular representation, their relative efficiencies, and how to assign credit or blame to choices among these alternatives upon unusual successes or failures. The motivations for employing multiple representations of extra-mental objects also apply to the representations of the mental objects figuring in self-knowledge. It is natural to identify these different views of objects as individual mental subagents, each with its own distinguished view of parts of the world, parts of the agent, or parts of both. This is roughly the position taken by MINSKY. However, if these different perspectives or subagents are to influence each other, they must be connected somehow, and the basic sorts of connections are those of reference and coreference. Both notions are necessary, for while one subagent might refer to another, intended extra-mental referents cannot be "grasped" in the same way, so at most the agent can intend that its representations of these objects share referents.

Unfortunately, the introduction of mutually knowledgeable and influential mental subagents into psychological theories poses many puzzling difficulties of formulation and interpretation. These difficulties appear most viciously in agents employing logical languages as systems of representation. Where classical logic and metamathematics usually seek ways of avoiding paradoxes of self-reference, the designers of artificial agents instead seem to seek them out. Fortunately, analysis of a narrow sort of self-knowledge (discussed presently) suggests a formal interpretation for these more widely self-referential systems, one which does not force us to accept any particular psychology for our agents, but instead allows similar formulation and exact comparison of the many variations we might think to explore.

Another sort of self-knowledge concerns the inferential relations between arbitrary representations instead of the economic relations between alternative coreferential representations. Many researchers have studied the uses of explicitly represented inferential relationships in constructing explanations, assigning blame for mistakes, and revising the agent's state of mind when its assumptions change. These inferential relationships need not be strictly deductive. While the most general use simply indicates what representations were computed from what, inferential records play a crucial role in so-called default reasoning. Default reasoning involves drawing conclusions in the absence of definite supporting or contrary evidence. Representations of the partial evidence for and the missing evidence against a conclusion permit the agent to make reasoned assumptions, "reasoned" in the sense that the agent can identify both the sources of the assumed conclusion and the specific information which indicates its retraction or reconsideration. The representations of inferential relationships describing reasoned assumptions also pose problems of interpretation, since the agent's drawing one conclusion may prevent it from drawing another. Fortunately, this problem has been solved, and below we extend the solution to handle the problem of multiple perspectives mentioned above.

A third important sort of self-knowledge concerns structural relationships between representations. The most studied structural relationship is that of *structure sharing*. Like the technique of multiple perspectives, structure sharing has economic motivations, namely minimizing the number of times one has to encode similar information and the amount of storage the agent must consume for the encodings. Like general inferential relationships, however, structure sharing need not entail coreference of the related representations. For example, the species of the cat family (lions, tigers, cheetahs, persians, etc.) may have no properties in common beyond those of mammals, since each cat species may lack some property shared by all other cat species. But to write down descriptions of each species is very tedious unless we write down a single description of a "prototypical" cat species (which we may choose to be one of the actual species) and describe every other species by its (presumably few) differences from the prototype. Since such family resemblances occur among the members of every natural kind, great economies can be realized in representing our common knowledge of the world. The most common sorts of structure sharing relations usually go by the names of "inheritance relations" and "virtual copies." As we demonstrate below, it is easy to interpret some of these structural relationships between subagents along with the previously mentioned ones.

§**3.** We first describe the mathematical framework in which we work, and then introduce the particular constitutive assumptions which characterize the three representational notions outlined above.

Our first fundamental assumption is that states of the agent can be decomposed into sets of mental elements or components. We write $\mathcal{D}$ to mean the *domain* of all possible mental components, so that if S is a state of the agent, then $S \subseteq \mathcal{D}$. Similarly, we assume that $\mathcal{A}$ is the set of all possible subagents, and that states of subagents may also be decomposed into elements of $\mathcal{D}$. For each $a \in \mathcal{A}$, we write $\mathcal{D}_a \subseteq \mathcal{D}$ to mean the subdomain of possible state-components of a, and require that every state component belong to at least one subagent, that is, $\mathcal{D} = \bigcup_{a \in \mathcal{A}} \mathcal{D}_a$. Thus if S is a state of the agent and $a \in \mathcal{A}$, then $S \cap \mathcal{D}_a$ is the state of subagent a. We do not require that every subagent exist in every state of the whole agent, that is, we allow $S \cap \mathcal{D}_a = \emptyset$.

Our second fundamental assumption is that every state component has an *interpretation* as a restriction on the possible states which contain it. Formally, we assume (and concretely define below) a function $\mathcal{I} : \mathcal{D} \to \mathbf{P}\,\mathbf{P}\,\mathcal{D}$ ($\mathbf{P}$ means power set), so that if $d \in \mathcal{D}$, then $\mathcal{I}(d) \subseteq \mathbf{P}\,\mathcal{D}$ is the set of potential states sanctioned by the element d. We encode in $\mathcal{I}$ the intended meanings of subagent state components for the relations standing between the state of the subagent and the states of other subagents. Note that these two assumptions permit several levels of decomposition of subagents into other subagents, as long as the interpretations chosen capture the intended synonymy of subagents with their subsocieties, for example by ruling out states in which one occurs without the other.

We define the *component-admissible sets* $\mathcal{Q} \subseteq \mathbf{P}\,\mathcal{D}$ to contain just the "self-satisfying" sets of state components. Formally,

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

That is, if $S \in \mathcal{Q}$, then all subagent states are in agreement as far as individual components of the state can tell. The third fundamental assumption of the framework is that every *admissible state* of the agent must be component-admissible. We write $\mathcal{S}$ for the set of admissible states of the agent, so our assumption is that $\mathcal{S} \subseteq \mathcal{Q}$. If $\mathcal{S} = \mathcal{Q}$, then all restrictions on admissible states are expressed in the "local" restrictions given by $\mathcal{I}$, and if $\mathcal{S} \neq \mathcal{Q}$, then there are "global" restrictions not expressed by $\mathcal{I}$. For example, $\emptyset \in \mathcal{Q}$ no matter how we choose $\mathcal{I}$, so nonemptiness of admissible states cannot be expressed as a local restriction.

These three fundamental assumptions exhaust the basic framework used in this paper. We now fill in the details of $\mathcal{D}$, $\mathcal{I}$, and $\mathcal{S}$ to characterize our particular mental society.

§**4.** While we do not require that subagents be completely representational, or that they employ any particular system of representation if they are completely representational, we do require a few minimal capabilities with which subagents can discuss each other. Our first particular constitutive assumption is that the state components of subagents can be further decomposed into "contents" indexed by the subagent. Formally, for each subagent $a \in \mathcal{A}$, we assume a set $\mathbf{C}_a$ such that $\mathcal{D}_a = \{a\} \times \mathbf{C}_a$. We further facilitate mutual reference by admitting subagents as possible contents, that is, $\mathcal{A} \subseteq \mathbf{C}_a$ for each $a \in \mathcal{A}$. To simplify matters, we assume that all content sets are the same set $\mathbf{C}$, and pretend that every content is a (possibly trivial) subagent by assuming $\mathcal{A} = \mathbf{C}$. These simplifications are innocuous since we can always rule out senseless elements by giving them the empty interpretation $\mathcal{I}(d) = \emptyset$ which prevents their inclusion in any admissible state. With these simplifications, we have $\mathcal{D} = \mathbf{C} \times \mathbf{C}$, and read $(a, b) \in \mathcal{D}$ as subagent a making the (possibly trivial) statement b. (We say "statement" here for want of a better term. Contents of subagents are statements only when $\mathbf{C}$ is a language, which we do not require.) For each state $S \in \mathcal{S}$, we find out what statements subagent a makes by means of the projection or *perspective* operator

$$p_a(S) = \{c \in \mathbf{C} \mid (a, c) \in S\}.$$

With this minimal notion of statements by subagents, we can describe the vocabularies of multiple perspectives, reasoned assumptions, and virtual copies. We introduce these vocabularies in turn by

means of abstract syntax functions. We also introduce separate interpretations $\mathcal{I}_1$, $\mathcal{I}_2$, and $\mathcal{I}_3$ for elements expressed in these vocabularies and define $\mathcal{I} = \mathcal{I}_1 \cap \mathcal{I}_2 \cap \mathcal{I}_3$, so that when more than one of these interpretations applies to a single element, their intersection is the complete interpretation of the element.

§**5.** The vocabulary of multiple perspectives is captured with three syntactic constructors on the set of contents. We assume the existence of functions $\big(,\big)$ [enlarged parentheses], $\langle , \rangle$, and $\Leftrightarrow$ from $\mathbf{C} \times \mathbf{C}$ into $\mathbf{C}$, so that for every $a, b \in \mathbf{C}$, we have $\big(a, b\big) \in \mathbf{C}$, $\langle a, b\rangle \in \mathbf{C}$, and $a \Leftrightarrow b \in \mathbf{C}$.

The $\big(,\big)$ constructor permits subagents to discuss the contents of other subagents, where we read $(a, \big(b, c\big)) \in \mathcal{D}$ as the statement made by a that subagent b makes the statement c. Since the constructor may be iterated, we can construct even more complex statements, such as

$$(c, \big(a, \big(c, \big(a, b\big)\big)\big)),$$

whose reading is left as an exercise.

The $\langle , \rangle$ constructor is the dual of $\big(,\big)$, and produces names for the multitude of relative perspectives. That is, we read $(\langle a, b\rangle, c)$ as the statement c made by the subagent corresponding to a's view of b. The corresponding reading exercise for this constructor is

$$(\langle\langle\langle c, a\rangle, c\rangle, a\rangle, b).$$

We make no assumptions of correctness or completeness about the "views" held by subagents about other subagents. That is, we allow an admissible state S to contain $(a, \big(b, c\big))$ even if $(b, c) \notin S$, and to contain (b, c) even if $(a, \big(b, c\big)) \notin S$. We leave pursuit of constitutive assumptions like correctness and completeness to future work. The only requirement we make is the intended connection between the dual constructor functions. This we express with the interpretation function $\mathcal{I}_1$ by requiring, for every $a, b, c \in \mathbf{C}$,

$$\mathcal{I}_1((a, \big(b, c\big))) = \{S \subseteq \mathcal{D} \mid (\langle a, b\rangle, c) \in S\}$$

and

$$\mathcal{I}_1((\langle a, b\rangle, c)) = \{S \subseteq \mathcal{D} \mid (a, \big(b, c\big)) \in S\}.$$

These interpretations have the consequence that

$$\mathcal{S} \subseteq \{S \subseteq \mathcal{D} \mid \forall a, b, c \in \mathbf{C} \quad (a, \big(b, c\big)) \in S \equiv (\langle a, b\rangle, c) \in S\},$$

which makes formulation of reasoned assumptions and virtual copies much easier.

Subagents use the $\Leftrightarrow$ constructor to specify coreferences. We read $(a, b \Leftrightarrow c)$ as a's thought that to it, b and c mean the same. Thus in a's view, every statement made by b will also be made by c, and vice versa. We capture this formally by defining, for every $a, b, c \in \mathbf{C}$,

$$\mathcal{I}_1((a, b \Leftrightarrow c)) = \{S \subseteq \mathcal{D} \mid p_{\langle a,b\rangle}(S) = p_{\langle a,c\rangle}(S)\}.$$

We complete the definition of $\mathcal{I}_1$ by defining $\mathcal{I}_1(e) = \mathbf{P}\,\mathcal{D}$ for every $e \in \mathcal{D}$ not covered above.

While the coreference constructor allows subagents to relate some of their own subperspectives, it cannot be used to relate "top-level" subagents. Since we require that every domain element belong to at least one subagent, every coreference statement must occur within some subagent, and hence only relate its subperspectives. That is, $(a, b \Leftrightarrow c)$ relates the perspectives of $\langle a, b\rangle$ and $\langle a, c\rangle$, not those of b and c. If our society is not to be a crowd of sleepwalkers, each unrelated to the others no matter how it dreams it is related, there must be connections between the subagents expressed either in $\mathcal{I}$ (which we do not do here) or in the $\langle , \rangle$ constructor. That is, we can read $\langle a, b\rangle = c$ as a's reference to c by means of b. Since the $\langle , \rangle$

constructor is defined, along with $\mathcal{D}$, independent of the element interpretations, all such references are "hard-wired" into the agent's realization, and cannot be changed by any action of the agent. We do not require that $\langle , \rangle$ be 1-1, and this allows us to "wire together" subagents by defining common references, for example by defining the constructor so that $\langle b, a\rangle = a = \langle c, a\rangle$, in which b and c can communicate and otherwise influence each other through a. If we imagine the human mind described in this way, $\langle , \rangle$ reflects the actual neural connections in the brain, while stated coreference relations using $\Leftrightarrow$ simply reflect the decisions of mental subagents.

We could of course introduce modifiable references by incorporating the $\langle , \rangle$ table into states. To do this, we need only redefine $\mathcal{D}$ to be $\mathbf{C}^2 \cup \mathbf{C}^3$, where elements $(a, b) \in \mathbf{C}^2$ are as before, and elements $(a, b, c) \in \mathbf{C}^3$ indicate entries in the constructor table, specifically, $(a, b, c) \in S$ means that $\langle a, b\rangle = c$ in S. We require that $\langle , \rangle$ be single-valued (but not necessarily complete) with the modified interpretation function

$$\mathcal{I}_1((a, b, c)) = \{S \subseteq \mathcal{D} \mid \forall d \in \mathbf{C} \quad (a, b, d) \in S \supset d = c\}.$$

We give subagents the capability to specify references by means of a constructor $\Rightarrow$ from $\mathbf{C}^2$ to $\mathbf{C}$, where $(a, b \Rightarrow c)$ is a's (ostensive) decision to use b to refer to c. This is formalized with the interpretation

$$\mathcal{I}_1((a, b \Rightarrow c)) = \{S \subseteq \mathcal{D} \mid (a, b, c) \in S\}.$$

Of course, we can get ostensive coreference from reference by using $(a, b \Rightarrow d)$ and $(a, c \Rightarrow d)$ instead of $(a, b \Leftrightarrow c)$. However, to keep the rest of the discussion as simple as possible, we forgo modifiable references for our original definitions, and leave recasting the subsequent definitions in terms of modifiable references as an easy exercise for the reader.

§**6.** The vocabulary of reasoned assumptions is captured with a single syntactic constructor of so-called finite simple reasons (see [Doyle 1982]). This constructor encodes each three finite subsets $A, B, C \subseteq \mathbf{C}$ as a single element of $\mathbf{C}$ written $A \setminus\setminus B \mid\!\!\!\vdash C$. We read $(d, A \setminus\setminus B \mid\!\!\!\vdash C) \in \mathcal{D}$ as d's commitment to make every statement in C whenever it also makes every statement in A and none of those in B. Formally, we define for every finite $A, B, C \subseteq \mathbf{C}$ and $d \in \mathbf{C}$

$$\mathcal{I}_2((d, A \setminus\setminus B \mid\!\!\!\vdash C)) = \{S \subseteq \mathcal{D} \mid A \subseteq p_d(S) \subseteq B^c \supset C \subseteq p_d(S)\},$$

and $\mathcal{I}_2(e) = \mathbf{P}\,\mathcal{D}$ for every other $e \in \mathcal{D}$. (B^c means the set-complement of B.) Note that elements of the form $(d, \emptyset \setminus\setminus \emptyset \mid\!\!\!\vdash A)$ correspond roughly to Minsky's "K-nodes." Combining simple reasons with mentioned perspectives allows phrasing versions of Stallman's [1981] "inter-theory inference rules," for instance

$$(d, \{(a_1, c_1)\} \setminus\setminus \{(a_2, c_2)\} \mid\!\!\!\vdash \{(a_3, c_3)\}).$$

§**7.** The vocabulary of virtual copies is captured in seven syntactic constructors, each of which represents an "indirect reference" version of the simple reason constructor. Specifically, we may substitute an indirect reference to the contents of a single subagent for one or more of the concrete sets mentioned by simple reasons. We indicate indirect reference to the contents of subagent a by the notation $@a$, so our constructors range from $@a \setminus\setminus B \mid\!\!\!\vdash C$ to $@a \setminus\setminus @b \mid\!\!\!\vdash @c$. The usual notion of simple structure sharing is then captured in elements like $(b, \emptyset \setminus\setminus \emptyset \mid\!\!\!\vdash @a)$, which we interpret as b's commitment to make every statement it thinks a makes. Formally, we define $\mathcal{I}_3$ so that

$$\mathcal{I}_3((d, @a \setminus\setminus B \mid\!\!\!\vdash C)) = \{S \subseteq \mathcal{D} \mid p_{\langle d,a\rangle}(S) \subseteq p_d(S) \subseteq B^c \supset C \subseteq p_d(S)\},$$

$$\vdots$$

$$\mathcal{I}_3(d, (@a \setminus\setminus @b \Vdash @c)) = \{S \subseteq \mathcal{D} \mid p_{\langle d,a\rangle}(S) \subseteq p_d(S) \subseteq (p_{\langle d,b\rangle}(S))^c \supset p_{\langle d,c\rangle}(S) \subseteq p_d(S)\},$$

and $\mathcal{I}_3(e) = \mathbf{P}\,\mathcal{D}$ for every other $e \in \mathcal{D}$.

§**8.** These definitions exhaust the vocabulary and interpretations of our simple society. We define $\mathcal{I}(d) = \mathcal{I}_1(d) \cap \mathcal{I}_2(d) \cap \mathcal{I}_3(d)$ for every $d \in \mathcal{D}$ and take $\mathcal{S} = \mathcal{Q}$. Although very brief, these constructions capture a surprisingly large part of the structure of important representational systems. We have left $\mathbf{C}$ unspecified, and some well-known representational systems can be captured largely as specific choices for $\mathbf{C}$. For example, if we choose $\mathbf{C}$ to be the set of all LISP S-expressions, the society bears close resemblance to SUSSMAN and STEELE'S constraint language. In this case, subperspectives correspond to the "parts" of constraints, or at the very deepest levels of embedding, the "values" of cells. The constraint language system also involves further restrictions on admissible states, such as closure under solutions to sets of equations between rational functions, but we avoid formalizing those here. As another example, if we choose $\mathbf{C}$ to be the set of sentences in a first-order logical language and adopt the modifiable reference definitions discussed earlier, the society bears close resemblance to the variant of WEYHRAUCH'S FOL system used in [DOYLE 1980]. In this case, subagents and perspectives correspond to "theories" and "subtheories," and $\Rightarrow$ corresponds to "semantic attachment." WEYHRAUCH'S system also incorporates a simplifier, evaluator, and automatic reflection mechanism, but we avoid pursuing these here as well.

§**9.** Our task is not yet done, however, for we have not yet formulated the "virtual" sense of virtual copies. If $S \in \mathcal{S}$, the above interpretations ensure that S contains all conclusions sanctioned by reasons and by structure sharing relations. That is, if $(b, \emptyset \setminus\setminus \emptyset \Vdash A) \in S$, then $A \subseteq p_b(S)$, and if $(b, \emptyset \setminus\setminus \emptyset \Vdash @a) \in S$, then $p_{\langle b,a\rangle}(S) \subseteq p_b(S)$. If the agent must realize all these elements in permanent storage, we have not achieved any economy of storage, even though we may have achieved economy in writing the information in the first place.

Similarly, admissible states contain all rewritings of all their elements in all equivalent perspectives. Since these agree in information, explicit realization in storage is uneconomical. One might also worry that admissible states must be infinite, but that is not so. Even if the constructor functions force $\mathcal{D}$ to be infinite, admissible states need not be infinite since we do not require that subagents be complete in their knowledge of each other. This means that most perspectives may be void, indeed, that finite states contain only finitely many nonempty perspectives. This situation is altered if we employ the sentences of a logical language for $\mathbf{C}$ and further require perspectives to be deductively closed, but we avoid those assumptions here.

We capture the motivations of virtual copies and virtual perspectives with the notions of extensions and admissible extensions. Suppose the agent only records some set $S \subseteq \mathcal{D}$ in storage. S need not be admissible itself if the agent interprets it as the "kernel" information from which to reconstruct a "full" admissible state. That is, if the agent needs to check the presence of some element in its state, it reconstructs the full state from S, checks for the element, remembers the answer, and then reclaims all storage except that used by S itself. We define $\mathit{Exts}(S)$, the *extensions* of S, by $\mathit{Exts}(S) = \{E \in \mathcal{S} \mid S \subseteq E\}$. We require that the full state reconstructed from S must be some $E \in \mathit{Exts}(S)$.

Unfortunately, extensions of S can contain, in addition to the missing elements virtually specified by S, elements completely unrelated to the kernel specifications. To see this, consider an analogous situation from logic. We may choose to economize storage in a logically structured agent by choosing and explicitly storing some axiomatization A of a deductively closed set S, that is, by picking A so that $S = \mathrm{Th}(A)$. In such an agent, deductively closed supersets correspond to extensions. But to reconstruct the initial set S from A, we cannot simply pick any deductively closed superset of A, since S is the smallest of these, and larger ones will contain extra unintended axioms and their consequences. To avoid the corresponding problem in our society, we introduce the notion of *admissible extension*. We say that

E is an admissible extension of S, written $E \in AExts(S)$, if and only if $E \in Exts(S)$ and E is *finitely grounded in* S. By finitely grounded we mean that for every $e \in E$ there is a finite sequence σ of elements of E (a "proof" of e in E) such that $e \in \sigma$ and every element of σ is either in S or is a required consequence of some elements appearing earlier in σ. Thus if $\sigma_i = (a, \big(b, c\big))$, either (1) $\sigma_i \in S$; or (2) for some $j < i$, $\sigma_j = (\langle a, b\rangle, c)$; or (3) for some $j < i$, $\sigma_j = (a, A \setminus\setminus B \mid\!\!\!\vdash C)$ where $\big(b, c\big) \in C$ and for every $d \in A$, (a, d) precedes σ_j in σ and for every $d \in B$, $(a, d) \notin E$; etc. We avoid presenting the full definition here, as it is not difficult to construct, merely tedious. A detailed development for the special case of finite simple reasons alone can be found in [Doyle 1982]. Since states contain reasoned assumptions, there may be several sets of assumptions possible in the kernel set S, and these can lead to distinct admissible extensions. For example, if just $(a, \emptyset \setminus\setminus \{b\} \mid\!\!\!\vdash \{c\})$ and $(a, \emptyset \setminus\setminus \{c\} \mid\!\!\!\vdash \{b\})$ are in S, then there will be two finitely grounded extensions, one containing (a, c) but not (a, b), and the other containing (a, b) but not (a, c).

While we here accept finitely grounded extensions as admissible extensions, they are inadequate to fully capture the usual notion of virtual copy. In current practice, it is crucial that successive queries agree, that is, that virtual information is conserved across reconstructions. But this cannot be guaranteed with multiple admissible extensions, since the agent might for one query construct E and next time construct $E' \neq E$, differing in some answers even though no kernel information has changed. Thus the virtual state is conserved only if the agent computes a unique admissible extension. Touretzky [1983] is currently developing restrictions on the sorts of information states can contain, restrictions designed to guarantee the existence and uniqueness of admissible extensions. He also motivates the aim of uniqueness by seeking parallel algorithms for reconstructing the virtual elements, and requiring that concurrent processes computing subsets of the virtual elements agree on their overlap. Touretzky's discoveries notwithstanding, I doubt that completely unoffensive restrictions on the vocabulary of the agent can alone guarantee uniqueness of finitely grounded extensions. I suspect that some applications demand a vocabulary sufficient to phrase ambiguities, and for these one appealing approach is to make the reconstruction algorithm, whether concurrently or serially realized, be a probabilistic algorithm. That is, when an ambiguity arises during reconstruction, the algorithm makes a random choice (random, not arbitrary). The intent of such deliberate randomization is to make every possible reconstruction equally likely or to occur with some specified frequency. If the agent wishes to judge its certainty on some question, it asks that question repeatedly. Questions with answers common to or absent from all admissible extensions never vary in their answer, while other queries exhibit uncertainty, waffling in response over time. If the alternative admissible extensions do not differ greatly, then most answers will be the same anyway no matter which admissible extension is chosen. [Doyle 1982] develops a theory of subjective probability by measuring the relative frequencies of different answers, but we cannot go into that here.

§**10.** Even if the ambiguities of admissible extensions are resolved, ambiguities due to multiple perspectives remain. Minsky and others have suggested that some abrupt changes in human behaviors and attitudes stem from changes in which subagent has control as "spokesman" over the communication or motor channels to the external world. In that view, there is no fixed notion of "self," but a different sense of self depending on which subagent gains control. One advantage of that view is that abrupt changes of attitudes are computationally trivial, for they stem from switching vantage points rather than from laborious revision of the state itself. The framework proposed here facilitates consideration of such proposals. For example, a natural problem is that of formulating precise notions of "abrupt" changes. If we decide when perspectives of different subagents are "similar," we can allow wide variations in which subagent is currently "self" as long as most of the self-image is conserved across self-changes, and single out as abrupt those changes of self which bring large or significant changes in the self-perspective. That is, if $p_a(S)$ and $p_b(S)$ are very similar, say if $p_a(S)$ and $p_b(S)$ differ by no more than 7 (± 2) elements, we might say that no major self-changes, only changes of attention, are involved in switches between a-self and b-self. Indeed, if the probabilistic approach to ambiguities of interpretation is adopted, then one need make no special provision for ambiguities due to self-changes. Can we develop measures of similarity on both states and perspectives

so that if S and S' are similar, so are $p_a(S)$ and $p_a(S')$, or vice versa? Unfortunately, we cannot pursue such questions here.

§**11.** There are many other possibilities to be explored in introducing notions of self into societies such as this. In [Doyle 1980] I advocated distinguishing a particular subagent, called ME, as the self. (I am less committed to that approach now.) When compared to the free-floating approach just discussed, the use of a fixed self-subagent appears to require significant computational costs for substantial perspective changes. (But see [McDermott 1982] and [Martins 1983].) In any event, distinguished perspectives still merit consideration, for it may be easier to endow them with limited completeness and correctness properties than amorphous agents. Specifically, one of the intents of my earlier proposal was to have the subagent ME be the authority on just what the state contained. The idea here is to construct the agent so that (modifiable references or not) $\langle \mathrm{ME}, a \rangle = a$ for every $a \in \mathbf{C}$ (including ME itself), if that is possible. I suspect it is not too difficult to achieve, and such organizations have obvious attractions for constructing agents possessing reflective powers. To pursue this idea, if one perspective admits the limited self-omniscience described above, does it follow that all do? That is, does global self-omniscience follow from local self-omniscience? I suspect not, but have no counterexample. It also seems certain that different perspectives can differ arbitrarily much even if both are mutually omniscient. If the subagents all use a logical language as a system of representation, well-known results indicate general limits to self-omniscience, but which sorts of limited self-knowledge can be introduced without difficulties arising? Kripke's analysis of truth indicates that even seemingly innocuous statements of mutual knowledge can in concert produce unreconcilable paradoxes. Since his theory involves a notion of groundedness resembling our notion of grounded extension, similar results seem likely here. Unfortunately, we must leave these questions for future study.

REFERENCES

Bobrow, D. G., and Winograd, T., 1977. An overview of KRL, a knowledge representation language, *Cognitive Science* **1**, 3-46.

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

Doyle, J., 1982. Some theories of reasoned assumptions: an essay in rational psychology, Pittsburgh: Department of Computer Science, Carnegie-Mellon University.

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

Doyle, J., 1983b. Admissible state semantics for representational systems, *IEEE Computer*, to appear.

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

Kripke, S. A., 1975. Outline of a theory of truth, *Journal of Philosophy* **72**, 690-716.

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

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

Minsky, M., 1965. Matter, mind, and models, *Proc. IFIP Congress*, 45-49.

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

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

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

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

Touretzky, D., 1983. Multiple inheritance and exceptions, Pittsburgh: Department of Computer Science, Carnegie-Mellon University.

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