Accounting for a Variety of Reasoning Data within a Cognitive Architecture

Ron Sun
Cognitive Sciences Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
rsun@rpi.edu

Xi Zhang
Department of CECS, University of Missouri, Columbia, MO 65211, USA

November 8, 2004

Abstract

This paper studies human everyday reasoning of a variety of forms. In particular, it explores the interplay among rule-based reasoning, (implicit) similarity-based reasoning, and (implicit) associative memory (intuition). In so doing, it incorporates both explicit and implicit forms of human reasoning in one unified framework, which is embodied in a cognitive architecture, CLARION.

Specifically, the paper first explores similarity in human everyday reasoning. A computational framework encompassing both rule-based and similarity-based reasoning provides explanations for human data. The simulation using CLARION demonstrates the role played by similarity-based reasoning in human everyday reasoning, and how such a reasoning process falls out of the structure of CLARION. Furthermore, this paper also explores the modeling of discovery tasks—tasks where sudden insights result from cumulative information, which is useful in better understanding reasoning involving intuition and insight. The situation is interpreted as involving mainly implicit associative memory with successive accumulation of information. The simulation within CLARION accurately captures some human data. Overall, the exploration of both similarity-based reasoning and intuition in this cognitive architecture leads toward a more comprehensive framework of human everyday reasoning.

1 Introduction

What is actual human everyday (i.e., mundane or "commonsense") reasoning like? Is it sufficiently captured by formal models developed by logicians and AI researchers, or is it somehow different? Is it completely explicit or is it mixed in terms of involving both explicit and implicit processes? Computationally speaking, what are the essential patterns in such reasoning?

A little background concerning the research described here is in order. Sun (1991) proposed a theory of human everyday reasoning based on a combination of rule-based reasoning and similarity-based reasoning, with a mixture of localist and distributed connectionist representations. This theory was further developed and elaborated in Sun (1995). The basic tenet of this theory is that, to a significant extent, human everyday reasoning consists of a combination of rule-based and similarity-based reasoning. Much of human everyday reasoning is reducible to these two types of processes. The intermixing of rule-based and similarity-based reasoning can lead to complex patterns of inferences as commonly observed in human everyday reasoning. And these two types of processes are captured within a unified connectionist model; that is, they fall out of the very same process (albeit with a combination of localist and distributed representations).

The theory was backed up by psychological evidence in the form of verbal protocols from Collins (1978) and Collins and Michalski (1989). These protocols were analyzed in Sun (1995) based on two mechanisms: rules and similarity (Tversky 1977, Hahn and Chater 1998). The analysis showed that the vast majority of the protocol data might be captured by the intermixing of these two mechanisms. This theory was crystallized into a two-component computational model whereby rule-based reasoning was carried out in one component with localist representation, and similarity-based reasoning in another with distributed representation (Sun 1995). Relevant to this approach, Sloman (1993) published a set of experiments, which provided support to the hypothesis of Sun (1991) (also Sun 1993, 1995). He found that similarity played a significant role in determining outcomes of inductive reasoning and similarity might be characterized by feature overlapping (as in Sun 1991). Later, Sloman (1998) described further experiments (in relation to category inclusion relations) that supported the hypothesis that there were two interacting mechanisms at work in human everyday reasoning (Sun 1993).

In the meantime, this theory (Sun 1995) evolved into a cognitive architecture, CLARION, which includes not only reasoning but also learning. It incorporates not only explicit representation (for explicit learning and explicit reasoning), but also implicit representation (for implicit learning and implicit reasoning). The interplay among rule-based reasoning, similarity-based reasoning, and intuition

(in the form of implicit associative memory) is also explored in this architecture.

In this paper, we attempt to analyze some data of human everyday reasoning in computational terms. We will instantiate our analysis in the form of computational models implemented in CLARION, which has been cognitively justified in various ways (Sun 2002) and therefore provides a cognitively grounded way of instantiating our analysis. First, as will be detailed later, we explore, within this architecture, the simulation of the categorical inference tasks in which rule-based reasoning and similarity-based reasoning intermix (Sun 1995, Sloman 1998). The simulation of the Sloman (1998) data accurately captures the human data. This simulation illustrates the respective role played by rule-based and similarity-based reasoning in human everyday reasoning. Furthermore, it demonstrates how such reasoning naturally falls out of CLARION. Thus, the simulation provides a detailed, precise, and process-based explanation of the data, avoiding the vagueness and the lack of specificity and precision that plague informal models and verbal theories. The advantage of computational simulation includes the fact that, due to the specificity and precision of computational models, different process-based explanations embodied by different simulations may be examined in detail and compared to each other precisely, and therefore a better understanding of the topic, or even a consensus, may be reached in the future.

Second, in this paper, we also explore another fundamental aspect of human everyday reasoning—intuition and insight. Specifically, we explore the discovery tasks, where sudden insights result from cumulative information (Bowers et al 1990, Schooler et al 1993, Metcalfe 1986). Such tasks are important for understanding finer details of human everyday reasoning (e.g., implicit processes), beyond explicit reasoning. We implement our analysis in CLARION, which provides a cognitively justified model (see Sun 2002) for describing the data in a precise and specific way (as opposed to the lack of precision and specificity in informal or generic theories). As will be detailed later, the simulation of the discovery task of Bowers et al (1990) within this architecture also accurately captures the human data. This simulation indicates the significant role played by implicit memory in human everyday reasoning.

It is worth noting that this work has been to some degree inspired by Barwise's various work (e.g., Barwise and Etchemendy 1995). For example, Barwise dealt with heterogeneous reasoning from the perspective of logic and mathematics, and focused, accordingly, on formal issues. From our perspective, reasoning may be heterogeneous at a more fundamental level beneath explicit representations that the mind consciously manipulates. Fundamentally, reasoning may be at once symbolic and non-symbolic. This point has been one of our underlying assumptions (Sun 1994).

In the remainder of this paper, we first describe the three pertinent experiments of Sloman (1998). We then describe the experiments of Bowers et al (1990). We present our interpretation of their experimental results. Next, we describe the (generic) cognitive architecture, CLARION, used in capturing human reasoning. Based on that, the particular setup of the architecture for capturing the experiments of Sloman (1998) is described. We then describe the results from simulating the experiments using CLARION. Similarly, we describe the setup within the architecture for capturing the data set of Bowers et al (1990), and the results of simulating their experiment. Finally, some general discussion completes the paper.

2 Empirical Data on Human Reasoning

Below, we will examine two sets of human data, concerning similarity and intuition in human reasoning, respectively.

2.1 The Categorical Inference Task

Let us examine some data that illustrate the interplay of similarity-based and rule-based reasoning (SBR and RBR, respectively) in human reasoning. We will look into the data from experiments 1, 2, 4, and 5 of Sloman (1998), which are most relevant to this issue.

In experiment 1, subjects were given pairs of arguments, each consisting of a premise statement and a conclusion statement. Some of these pairs of arguments may be in the form of *premise specificity*:

a. All flowers are susceptible to thrips. $\Longrightarrow$ All roses are susceptible to thrips.

b. All plants are susceptible to thrips. $\Longrightarrow$ All roses are susceptible to thrips.

Some other pairs of arguments may be in the form of *inclusion similarity*:

a. All plants contain bryophytes. $\Longrightarrow$ All flowers contain bryophytes.

b. All plants contain bryophytes. $\Longrightarrow$ All mosses contain bryophytes.

Subjects were to pick the stronger of the two arguments from each pair. 73 subjects were tested and each was given 18 pairs of arguments (among other things not related to this task). [1]

[1] These arguments may be viewed as enthymematic. But they are more than just enthymemes, due to the use of similarity-based reasoning, as will be shown later.

The results showed that the more similar argument from each pair of arguments was chosen 82% of times (for inclusion similarity) and 91% of times (for premise specificity). t tests showed that these percentages were significantly above chance, either by subjects ($t(72) = 18.64$ and $t(72) = 33.09$ for premise specificity and inclusion similarity, respectively; $p < 0.0001$) or by argument pairs ($t(8) = 6.97$ and $t(8) = 15.61$ respectively; $p < 0.0001$).

It should be apparent that if only RBR (e.g., based on logics) was used, then similarity should not have made a difference, because the conclusion category was contained in the premise category and thus both arguments in each pair should have been equally, perfectly strong. Therefore, the data suggested that SBR (as distinct from RBR or logics capturing category inclusion relations) was involved to a significant extent.

In experiment 2, subjects were instead asked to rate the likelihood ("conditional probability") of each argument. Ratings could range from 0 to 1. 18 subjects were tested.

The mean rating was 0.89 for inclusion similarity and 0.86 for premise specificity. Both were significantly below 1, both by subjects ($t(17) = 2.75$ and $t(17) = 3.23$ respectively; $p < 0.01$), and by arguments ($t(17) = 8.87$ and $t(17) = 6.14$ respectively; $p < 0.0001$). Again, we notice that it would have been the case that the outcome was 1 if only RBR had been used (because the conclusion category was contained in the premise category). Thus, SBR was significantly present here too. Indeed, ANOVA showed that across subjects, there was a significant main effect of similarity (low vs. high; $F(1, 17) = 18.90, p < 0.001$). So was the case across argument pairs ($F(1, 16) = 12.64, p < 0.001$).

In experiment 4, subjects were asked to rate the likelihood of each argument. Ratings could range from 0 to 1. However, in this case, each category inclusion relation was specifically presented as part of each argument. For example,

All plants contain bryophytes. All mosses are plants. $\Longrightarrow$ All mosses contain bryophytes.

The results showed that the mean judgment was 0.99. 23 out of 27 subjects gave all 1's. 32 out of 36 arguments received judgments of all 1's (excluding one individual who gave 0.99 throughout). In other words, the similarity-based phenomena almost disappeared. Instead, it appeared that an explicit RBR mode based on category inclusion relations was used.

Experiment 5 was similar to experiment 2, in that ratings were obtained. However, before any rating was done, subjects were asked to make category inclusion decisions. Thus, in this case, subjects were reminded of rule-based reasoning explicitly involving category inclusion relations. Therefore, they

were more likely to use RBR, although probably not as much as in experiment 4, due to the separation of category inclusion judgment and argument likelihood rating in the experiment procedure (unlike that of experiment 4).

The results showed that no one of the 18 subjects gave a likelihood judgment of 1 for every argument, indicating that SBR might be at work. Compared with experiment 2, having subjects make category inclusion judgments increased the likelihood rating. The mean judgment for experiment 5 was 0.92 as opposed to 0.87 for experiment 2. [2] This increase appeared to reflect the increased involvement of RBR. Nevertheless, ANOVA showed a significant effect of similarity (low vs. high), across subjects ($F(1,17) = 9.33, p < 0.01$), and across argument pairs ($F(1,16) = 11.42, p < 0.01$).

It is important to note that, given the co-existence of RBR and SBR, conventional logics (or their psychological variations; e.g., Rips 1994) may be used to capture the RBR aspect of the human data, but not the SBR aspect, because they would not be able to tell the difference between the two arguments within each pair in terms of their strength difference. Although one may argue that logics could encode whatever similarity relationships humans employ in SBR (for capturing the evaluation of strengths), such a "solution" would not be satisfactory for many reasons (including its ad hoc nature and its high representational costs; more later).

2.2 Interpretation of the Categorical Inference Data

Based on the analysis above, we see that varying proportions of RBR and SBR were involved. Among them, experiment 1 and experiment 2 both involved SBR to a very significant extent. Experiment 4 involved explicit use of categorical relations, and thus mainly RBR. Experiment 5 involved more of SBR, along with RBR.

To account for the data mechanistically (computationally), let us look into the process involved. Mechanistically, we envisage the following process for dealing with the premise and the conclusion statement of each argument: First, a premise statement is presented. Each premise statement (e.g., "all flowers are susceptible to thrips") is encoded as a rule in the system. The two concepts involved in it are encoded as well, both implicitly (in distributed representations) and explicitly (as individual units). Inclusion relations, such as "roses are flowers", are already existent in the system and encoded also as rules. When it comes to dealing with the conclusion statement (e.g., "all roses are susceptible to

[2] However, the difference was not statistically significant by subjects, although significant by arguments ($t(35) = 3.81, p < 0.0001$).

thrips"), the first concept of the conclusion statement (e.g., "rose") is presented. Due to the similarity between the first concept of the conclusion statement and the first concept of the premise statement, the encoded rule representing the premise statement is partially activated. As a result, the target concept (the second concept of the premise statement; e.g., "thrips") is also partially activated, the extent of which is proportional to the afore-mentioned similarity. [3]

When there is a pair of such an argument, this process is repeated. The two partial activations of the target concept are temporally stored. Then the two partial activations are retrieved and compared. In case of forced choice, one of them is selected. The selection favors the more strongly activated one (but the process could certainly be noisy).

Of course, the process described above is under the direction (control) of some executive functions. Since such executive control is of the usual variety, we will not analyze it in any more detail here.

Later in the paper, we will simulate this task of categorical inference (1) for validating this analysis and (2) for further testing the CLARION cognitive architecture. The simulation shows indications of the significance of similarity-based reasoning, as opposed to rule-based/logical reasoning or probabilistic/Bayesian reasoning (cf. Anderson and Lebiere 1998). The significant role of similarity-based reasoning here distinguishes this type of reasoning from the kind of reasoning captured naturally by usual production systems (see more discussions of this point in section 5).

2.3 The Discovery Task

Let us now turn to examine some human data that illustrates intuition and insight in human everyday reasoning (Tversky and Kahneman 1983, Sun 1991). Specifically, we will look into the data from Bowers et al (1990), focusing on their third experiment, which shows that intuition and sudden insight may be underlaid by a gradual "warming up" process (first warming up to a hunch and later to a conviction).

First, some clarification is needed. Although intuition has been defined as "the immediate apprehension of an object by the mind without the intervention of any reasoning process" (Oxford English Dictionary), or "immediate knowledge, as in perception or consciousness, distinguished from mediate knowledge as in reasoning" (Webster's Dictionary), we instead view intuition as a kind of reasoning (Sun 2002). Reasoning encompasses both explicit processes (explicit rules and logics) on the one hand,

[3] The heavy involvement of similarity here distinguishes this type of reasoning from the kind of reasoning captured by common production systems (Sun 1994).

and implicit processes (intuition and insight) on the other (Sun 1991, 1995). In fact, intuition and insight are important components of human reasoning. They supplement and guide explicit reasoning. The latter has been amply documented in cognitive science and AI (see, e.g., Collins and Michalski 1989, Davis 1990, Yang and Johnson-Laird 2001), but not yet the former. Thus, more studies of everyday reasoning involving intuition and insight are needed.

In the discovery task of Bowers et al (1990), during each trial, a set of 15 clue words (that is, an "item") was presented to subjects, one word at a time. Each clue word was a response to a stimulus word in the Kent-Rosanoff word association test (Kent and Rosanoff 1910). The first 12 clue words occurred 5 or less times out of 1000 as a response to the stimulus word, and they were randomly assigned to position 1-12. The last three clue words occurred more than 5 times and were randomly assigned to position 13-15. Subjects' task was to identify the word that was associated with each of the 15 clue words.

The clue words from each set were presented one at a time. Subjects were required to generate a word to which each of the clue words was associated after the presentation of each clue word. They were given 10-15 seconds after the presentation of each clue word. If subjects viewed a generated word as a potential solution, they were to check-mark it (indicating a "hunch"). When they were convinced that the word was a solution, they were to mark it with an X (indicating a "conviction").

As reported in Bowers et al (1990), in a sample of 100 subjects, subjects arrived at a hunch on about the 10th clue word on average (M=10.12, SD=4.55). The average number of clue words it took to go from a hunch to a conviction was about 1.79 (SD=0.96).

2.4 The Interpretation of the Discovery Task Data

As suggested by Bowers et al (1990), subjects could respond discriminatingly to coherence that they could not explicitly identify, and that this implicit recognition of coherence guided subjects gradually toward an explicit representation of a hunch. Subjects "warmed up" to the solution in an incremental and gradual manner. That is, whereas hunches or convictions might surface quite suddenly into consciousness, implicit cognitive processes were rather continuous. An implicit representation gradually gained strength. When the level of activation reached a certain degree (crossing a certain threshold), the implicit representation triggered an explicit one. (On the other hand, an explicit representation might also be activated as a result of relevant explicit knowledge, e.g., explicit rules. However, there were often very few such explicit rules, and therefore they were often irrelevant in determining the

outcomes, as indicated by the initial implicitness of recognition and by the gradual explicitation shown by human subjects in this task.) A "hunch" was indicated as a result of an emerged and sufficiently activated explicit representation. Even after that point, its strength might continue to grow, and thus eventually, subjects might indicate a "conviction"—an even stronger explicit representation.

Mechanistically (i.e., computationally), we can easily imagine that people are frequently "trained", deliberately or incidentally, with word associations (for example, desk-chair, pen-paper, and so on) in everyday life. Such training forms associations between pairs of words with varying degrees of strengths, based in part on frequencies of co-occurrences. Association formation happens mainly in the implicit memory (specifically, in the part of the implicit memory that is not concerned with procedural, or action-centered, information), because of the (mainly) incidental nature of the training scenario. At the test setting of the experiment of Bowers et al (1990), clue words are presented one at a time. At the presentation of each clue word, all associated words are activated to certain extents (some strongly while others weakly). After the presentation of each clue word, more activations are accrued to the associated words. Gradually, the activations of some words in the implicit memory become stronger and stronger. As a result of the implicit memory (as well as explicit rules possibly), explicit representations are activated in the explicit memory (specifically, the part of the explicit memory that is concerned with the general non-action-centered knowledge about the world). Eventually, a threshold (the threshold for hunches) is crossed, and thus a hunch is found. Furthermore, when more clue words are presented, more activations are accrued to the explicit representations. A second threshold (the threshold for convictions) is crossed, and thus a conviction is declared.

Note that the same as in the previous task, the process described above is under the control of some executive decision making functions. Such control is of the usual variety, and thus we will not analyze it further.

In the remainder of this paper, we will attempt to implement the above analysis of the processes underlying the two tasks in computational models, which serve to instantiate, substantiate, verify, and validate the analysis. As mentioned earlier, the simulations are conducted within CLARION.

3 The CLARION Model

3.1 Overall Structure

CLARION is an integrative cognitive architecture with a dual representational structure (Sun et al

2001, Sun 2002). That is, it consists of two levels: the top level captures *explicit* processes and the bottom level *implicit* processes. See Figure 1.

INSERT FIGURE 1 ABOUT HERE

First, the inaccessible nature of implicit knowledge is suitably captured by subsymbolic distributed representation provided by a backpropagation network. This is because representational units in a distributed representation are capable of accomplishing tasks but are subsymbolic and generally not individually meaningful (see Smolensky 1988, Sun 1995). This characteristic of distributed representation accords well with the (direct) inaccessibility of implicit knowledge.

In contrast, explicit knowledge may be captured in computational modeling by a symbolic or localist representation (Clark and Karmiloff-Smith 1993), in which each unit is easily interpretable and has a clear conceptual meaning. This characteristic captures the property of explicit knowledge being (directly) accessible and manipulable (Smolensky 1988, Sun 1995). [4]

This radical difference in the representations of the two types of knowledge leads to a two-level model whereby each level using one kind of representation captures one corresponding type of process, either implicit or explicit. The model may select to use one level or the other, based on current circumstances (e.g., experimental conditions; see Sun 2002 for details). When both levels are used, the outcome from the two levels may be integrated in some ways (which may be partially domain specific; Sun 2002).

At each level of the model, there may be multiple modules, both *action-centered* modules and *non-action-centered* modules (Schacter 1990, Moscovitch and Umilta 1991). The reason for having both action-centered and non-action-centered modules (at each level) is because, as it should be obvious, action-centered knowledge (roughly, procedural knowledge) is not necessarily inaccessible (directly), and non-action-centered knowledge (roughly, declarative knowledge) is not necessarily accessible (directly). Although it was argued by some that all procedural knowledge is inaccessible directly and all

[4]Note that explicitness/implicitness has a lot to do with ease of access, regardless of mechanisms or processes of access (Kirsh 1990). It may be accounted for, for example, by a theoretical/mathematical measure known as Kolmogorov complexity (Sun 2002). This is the justification behind mapping implicit/explicit knowledge to distributed/localist representation. For more discussions of why distributed representation is more suitable for capturing implicit knowledge and localist one for explicit knowledge, see Sun (2002). Since this is not a philosophical paper, we will not repeat the detailed arguments here.

declarative knowledge is directly accessible, such a clean mapping of the two dichotomies is untenable in our view. Henceforth, we will refer to these two sets of modules as the *action-centered subsystem* (or the ACS for short) and the *non-action-centered subsystem* (or the NACS for short), respectively. There are also some other components, such as a goal structure (GS), working memory (WM), episodic memory (EM), and so on, details of which are not important to this work and will not be described here.

3.2 The Non-Action-Centered Subsystem

In this work, we will focus on the NACS, due to the declarative nature of the task. This subsystem, as stated earlier, consists of (1) a top level, which is made up of a set of explicit associative rules, and (2) a bottom level, which is made up of implicit associative memories (Sun 2002).

The top level of the NACS is termed the general knowledge store (or the GKS for short). In the GKS, the essential elements are *chunks*, each of which is specified by a set of dimension-value pairs (i.e., attribute-value pairs) that describes an entity, along with a chunk label. Each chunk is represented by a chunk node at the top level. The dimensional values of a chunk are represented individually at the bottom level. Thus a chunk node is linked to the corresponding dimension-value nodes at the bottom level. In addition, links between chunks, termed *associative rules*, encodes explicit associations between pairs of chunks. These rules may be learned from externally provided information or information from the bottom level (i.e., from implicit memory).

When there are multiple rules reaching the same conclusion, they are combined in the strength of the conclusion chunk as follows:

$$S^c_{c_i} = \max_{\text{all rules } j \text{ leading to } c_i} S^a_j$$

where c_i indicates the ith chunk at the top level, and

$$S^a_j = \sum_k A_k \times W_k$$

where A_k is the strength of the kth chunk in the condition of the rule (which is 1 if the chunk is externally given), W_k is the weight of the kth chunk in the condition of the rule (the default is $W_k = 1/n$, where n is the number of chunks in the condition of the rule).

There are two relevant parameters here. A density parameter (d_c) determines the minimum frequency of invocation (encoding, re-encoding, extraction, re-extraction, or activation) necessary in

order to keep an existing chunk. For example, if $d_c = 1/100$, then there should be at least one invocation (encoding, re-encoding, extraction, re-extraction, or activation) of a chunk for every 100 steps in order to keep it. Similarly, another density parameter (d_a) determines the minimum frequency of invocation (encoding, re-encoding, extraction, re-extraction, or application) necessary in order to keep an existing associative rule. [5]

Next, at the bottom level of the NACS, associative memory networks (or AMNs for short) encode implicit associations. Specifically, among many other possibilities, a backpropagation network can be used for implementing this memory. Associations are formed by mapping an input to an output. For example, the network may learn to hetero-associate a pattern with another: a pattern is presented as the input, and another (different) pattern to be associated with it is presented as the desired output. Through repeated training, the network may be able to output the desired pattern when a corresponding input pattern is presented. This is useful for, for example, establishing connections between concepts (such as linking climate information of a region to its agricultural products, or linking a person to his profession).

Finally, the integration of outcomes from the two levels (the GKS and the AMNs) is simple: When the output of the AMNs is activated, the chunk in the GKS corresponding to the output from the AMNs will be activated to the same extent. The activations from the AMNs and the GKS may be combined in various ways. The default way is through a *max* function.

Note also that all of the operations of the non-action-centered subsystem, including the operations of both levels, are under the direct control of the action-centered subsystem, which makes action decisions each step of the way. To do so, the top level of the ACS consists of a set of explicit action rules, either externally given or extracted from the bottom level (from implicit knowledge), while the bottom level consists of implicit decision networks (trained with reinforcement learning algorithms, negligible in this task). For details regarding the ACS and its parameters, see Sun et al (2001) and Sun (2002). We will not get into its details here, as they are not directly relevant to the simulations in this paper.

3.3 Similarity

Furthermore, similarity-based reasoning falls out of encoding with chunk nodes and dimension-value nodes. The general idea is as follows: A known (given or inferred) chunk is compared with another

[5] The default is $d_a = d_c = 1/100$.

chunk. If their similarity is high enough, then the other chunk is inferred. The strength of a chunk c_j as the result of similarity-based reasoning is:

$$S_{c_j}^{c,s} = \max_i (S_{c_i \sim c_j} \times S_{c_i}^c)$$

where $S_{c_i \sim c_j}$ measures the similarity from c_i to c_j (Tversky 1977), $S_{c_i \sim c_j} \times S_{c_i}^c$ measures the support to c_j from the similarity, and i ranges over all the chunks. The (default) similarity measure (Sun 1995) is: [6]

$$S_{c_i \sim c_j} = \frac{N_{c_i \cap c_j}}{f(N_{c_j})}$$

where $S_{c_i \sim c_j}$ denotes the similarity from c_i to c_j. In the formula, $N_{c_i \cap c_j}$ is the weighted sum of all the identically valued dimensions of c_i and c_j (among all the specified dimensions of c_j), that is, $N_{c_i \cap c_j} = \sum_{k \in c_j \cap c_i} V_k^{c_j} \times D_k$, where D_k in this case represents dimension k in chunk c_j and has a value of 1. The weights $V_k^{c_j}$ in the above weighted sum are specified with respect to c_j (the target of similarity, not the source of it). The default is that these weights are the same and equal to 1. On the other hand, N_{c_j} is the weighted sum of the specified dimensions of c_j; that is, $N_{c_j} = \sum_{k \in c_j} V_k^{c_j} \times D_k$, where $D_k = 1$ and $V_k^{c_j} = 1$ (by default). f is a superlinear, but close to linear, function, such as $f(x) = x^{1.10}$. Thus, the similarity measure is limited to $[0, 1)$. Clearly, similarity is computed based on feature overlapping (i.e., dimension-value overlapping): The more similar two concepts are, the more feature overlapping there is between the two, and vice versa. For further details, see Sun (1995).

From the above abstract specification of SBR, it is natural that similarity-based reasoning be carried out with a combination of the top level and the bottom level. Furthermore, similarity may be automatically computed whenever reasoning involves multiple chunks that are similar to one another. Therefore, there is no dedicated representation of similarity between any two chunks. The afore-specified abstract SBR formulation is carried out within the structure of CLARION. The weights for bottom-up activation of a chunk (i.e., for the activation of a chunk node from activated dimension-value nodes at the bottom level) are:

$$W_k^{c_j} = \frac{V_k^{c_j}}{f(\sum_{k \in c_j} V_k^{c_j} \times D_k)} \tag{1}$$

where $W_k^{c_j}$ is the weight of the kth dimension-value node of chunk j (at the bottom level), and $V_k^{c_j}$ and D_k are as specified before. Thus, in turn, the bottom-up activation $S_{c_j}^{c,s}$ is determined as follows:

$$S_{c_j}^{c,s} = \sum_{k \in c_j} W_k^{c_j} \times A_k = \sum_{k \in c_j} \frac{V_k^{c_j} \times A_k}{f(\sum_{k \in c_j} V_k^{c_j} \times D_k)} \tag{2}$$

[6] Note that there are many other possible similarity measures. See, for example, Tversky (1977) or Sun (1995) for some of them. Compared with these options, the default option described above is preferable (Sun 1995).

where A_k is the activation of the dimension-value node k of chunk c_j (at the bottom level). This process implements SBR as abstractly specified earlier (Sun 1994). [7] Thus, it requires no additional mechanism (process or representation).

Similarity-based and rule-based reasoning can be inter-mixed. When both SBR and RBR are employed, we have:

$$S^c_{c_i} = \max(\alpha \times \max_{j:\text{all rules leading to } i} S^a_j,$$
$$\beta \times \max_{j:\text{all chunks similar to } c_i} (S_{c_j \sim c_i} \times S^c_{c_j}))$$

where α and β are two constants that balance the two measures (rule versus similarity), [8] and $S_{c_j \sim c_i}$ is the similarity measure. As a result of mixing SBR and RBR, complex patterns of reasoning can emerge. As explicated in Sun (1995), the conclusion from one step of reasoning can be used as the starting point of the next step. The iterative process of combined rule-based and similarity-based reasoning allows all possible conclusions to be reached (including "inheritance" reasoning; see Sun 1995). These different sequences together capture essential patterns of human everyday reasoning (see Sun 1995 for details).

3.4 Previous Simulations

It is worth noting that CLARION has been successful in simulating a variety of cognitive tasks. These tasks include serial reaction time tasks, artificial grammar learning tasks, process control tasks, alphabetical arithmetic tasks, and the Tower of Hanoi task (Sun 2002). In addition, we have done extensive work on a complex minefield navigation task (Sun et al 2001, Sun and Peterson 1998). Thus, the cognitive validity of CLARION is established to a certain extent. We are now in a good position to extend the effort to the capturing of a wide range of human reasoning and memory processes, through simulating reasoning and memory data. This paper is but one aspect of this effort.

[7] The minor difference from the abstract specification (with the default similarity measure) lies in the use of A_k (as opposed to D_k times $S^c_{c_i}$ as in the abstract specification).

[8] α and β can be reduced to other existing parameters—rule weights and similarity measures. We present them here as separate parameters in order to highlight some aspects of the simulation setup and facilitate the discussion of the simulations later on.

4 Simulations of Human Reasoning

4.1 Simulation Setup of Categorical Inference

As explained before, there were a number of different experimental settings in the categorical inference task. For simulating these experimental settings, the following manipulations were used: For simulating settings where SBR was dominant, RBR was de-emphasized. For simulating settings where RBR was dominant, RBR was emphasized. The relative emphasis of the two methods (RBR versus SBR) was accomplished through the *balancing* parameters. We set $\alpha = 0.5$ and $\beta = 1.0$ for experiments 1 and 2, because of the heavy reliance on SBR as opposed to RBR, as suggested by the analysis of the human data (see the earlier discussion of the human data). For simulating experiment 4, they were set at $\alpha = 1.0, \beta = 1.0$, because this setting prompted more reliance on RBR as indicated by the human data. For simulating experiment 5, they were set at $\alpha = 0.88, \beta = 1.0$, because the experiment involved an intermediate level of reliance on RBR as suggested by the human data. In all, these values were set in accordance with our interpretations of what happened under the different experimental conditions respectively.

For simulating the task, through training, at the top level of the NACS (the GKS), all relevant category inclusion relations, such as "flowers are plants" or "mosses are plants", were encoded as associative rules. Through training, chunks were used to represent concepts such as "flowers" and "plants". The dimensional values of these concepts were represented in the AMN, and the chunk nodes representing these concepts in the GKS were linked to the AMN representation.

In the AMN, although associative memories were present, they were not very relevant for the performance of this task, because there was no sufficient prior training of the network with any data directly relevant to this task (because this task used unknown or made-up concepts, which were presented only once). [9] The integration of the outcomes of the two levels was through a *max* function. The real effect of integrating the outcomes of the two levels amounted only to adding a small amount of noise to the outcomes of the GKS.

Training of the model, before the simulation of the experimental test of Sloman (1998), consisted of presenting categorical features (dimension-value pairs) along with the category labels, to both levels

[9] For the associative memory network, the number of input units was 1800 (for representing all chunks specifiable with 60 dimensions of 30 possible values each), the number of hidden units was 500, and the number of output units was 1800. The learning rate was 0.3 and the momentum was 0.1.

of the NACS. Note that repeated presentations were not required. The one-pass presentation enabled the formation of chunks and associative rules in the GKS, but not much implicit knowledge in the AMN.

During the test, when a category name was given, the category name was matched with a corresponding chunk label. The matching chunk was activated to the full extent (i.e., 1). Then, through associative rules as well as similarity-based processes, conclusion chunks were also activated (to varying extents), combining SBR and RBR according to the balancing parameters. Conclusion chunks were retrieved along with their strengths.

For simulating ratings of conclusions (as in experiments 2, 4, and 5), the strengths of chunks derived from a proper combination of the results of SBR and RBR (as determined by the balancing parameters) were directly used. However, for simulating forced choices (as in experiment 1), a stochastic decision process based on the Boltzmann distribution was used to select between two possible outcomes.

The following action rules (among many other action rules) were implemented in the ACS of CLARION for directing the performance of the NACS in this task:

> If goal= forced-choice-task and no source category has been presented, then present the first source category, obtain the rating of the target category, and store it in the working memory.
>
> If goal= forced-choice-task and one source category has been presented, then present the second source category, obtain the rating of the target category, and store it in the working memory
>
> If goal= forced-choice-task and both source categories have been presented, then conduct a stochastic selection from the two ratings in the working memory, and report the result.
>
> If goal= rating-task and no source category has been presented, then present the first source category, obtain the rating of the target category, and store it in the working memory.
>
> If goal= rating-task and one source category has been presented, then present the second source category, obtain the rating of the target category, and store it in the working memory.
>
> If goal= rating-task and both source categories have been presented, then report the two ratings from the working memory.

These rules were "fixed" rules for this task, presumably derived from a priori knowledge and task

instructions (given to subjects prior to experiments). The goals involved in these rules were set in the goal structure, when the task instructions were given before the test began.

4.2 Simulation Results of Categorical Inference

We simulated the data from experiments 1, 2, 4, and 5 of Sloman (1998) as described earlier. For each experiment, a set of simulation runs (i.e., simulated "subjects") equal to the number of the human subjects involved in the original human experiments were used. The results and the statistical analysis of the results were as follows.

As described before, in experiment 1, subjects were to pick the stronger of the two arguments from each pair. The simulation of experiment 1 showed, the same as the human data, that the more similar argument from each pair of arguments was chosen more often: 82% of times (for inclusion similarity) and 83% of times (for premise specificity). t tests showed that these percentages were significantly above chance, either by subjects ($p < 0.001$) or by argument pairs ($p < 0.001$), the same as in the human data. In our simulation setup, there was a significant involvement of SBR (with $\alpha = 0.5, \beta = 1.0$). If only RBR had been used, then similarity could not have made a difference, and thus both arguments in a pair should have been equally strong. This simulation demonstrated that the significant involvement of SBR (as distinct from RBR) in producing the human data of this experiment was a reasonable interpretation (see the earlier exposition of the human experiments), given the close match with the human data.

In experiment 2, subjects were instead asked to rate the likelihood of each argument. In this simulation, the mean rating was 0.86 for inclusion similarity and 0.87 for premise specificity. Both were significantly below 1, different from what would have been predicted if only RBR had been used, both by subjects ($p < 0.001$) and by arguments ($p < 0.001$), the same as in the human data. ANOVA also showed that across subjects and across argument pairs, there was a significant main effect of similarity (low vs. high; $p < 0.001$). With the same setup as the previous simulation, this simulation again demonstrated the same pattern of significant involvement of SBR as in the human data (which could not be naturally captured by usual RBR or logics).

In experiment 4, subjects were asked to rate the likelihood of each argument, right after being presented corresponding category inclusion relations. The simulation produced the mean judgment 0.99, the same as the human data. Compared with experiment 2, explicit RBR based on category inclusion was much more prominent in this case, as specified in our simulation setup ($\alpha = 1.0, \beta = 1.0$),

which captured the human data accurately.

In experiment 5, ratings were obtained after subjects were asked to make category inclusion decisions. In this case, subjects were reminded of RBR involving category inclusion relations and therefore they were more likely to use RBR (compared with experiment 2), although not as much as in experiment 4. In the simulation, the mean judgment for experiment 5 was 0.91 for both inclusion similarity and premise specificity, as opposed to 0.86 and 0.87 for the two cases in experiment 2. ANOVA also showed a significant main effect of similarity (low vs. high), across subjects ($p < 0.001$), and across argument pairs ($p < 0.001$). This simulation replicated the human data well, which showed that our interpretation as embodied in the simulation setup ($\alpha = 0.88, \beta = 1.0$), that is, less involvement of RBR compared with experiment 4 but more compared with experiment 2, was a reasonable one.

In all, the simulation of this task successfully substantiated and thereby validated our interpretation and analysis of human performance in this task as described earlier. In particular, mechanistic processes formulated on the basis of CLARION were carried out through the simulation, which replicated the human data. Due to the close match with the human data, these formulated mechanistic processes constitute a detailed theoretical explanation and a process-based hypothesis regarding human reasoning. Such a detailed explanation/hypothesis would not be possible, without detailed computational simulations.

4.3 Simulation Setup of the Discovery Task

Let us now turn to the simulation setup for the discovery task. In simulating this task, the model was first trained. Different from the previous simulation, in this case, much longer training was required to capture gradually formed implicit word associations that subjects possessed. During the training of the model, pairs of words were presented to the AMN (as well as the GKS). The input to the AMN included three components: the current clue word, the working memory content, and the current goal. The input nodes of the AMN corresponded to dimension-value representations of words and goals. The output nodes of the AMN corresponded to the dimension-value representation of a word. [10] Each of the first 12 words on a list in the stimulus material was used for training, paired with the target word. Each of these words was used about 4% of the times, for a total of about 48% of the training time. Each of the 12 words was presented as the input to the AMN and the target word as

[10] In fact, the input and output should be phonological and morphological features of words. However, for the sake of simplifying the simulation, we used artificially constructed features. This simplification did not affect the outcome of the simulation.

the desired output for the AMN. [11] These associations were somewhat under-trained (thus the AMN did not perform well at the end of the training), capturing weak, implicit associations between these pairs of words. Each of the last 3 words on a list in the stimulus material was also used for training, again paired with the target word. Each of these words was used for training 17% of the times, for a total of 51% of the training time. A total of 10 lists of words (10 training "items") was used. [12]

Each word was represented as a chunk node in the GKS, due to the presentation during training of these words as input and output. The dimensional values of these chunks were represented in the AMN, and thus chunk nodes were linked to the AMN. However, due to the relatively infrequent presentation of association pairs, there was the encoding of relatively few explicit associative rules in the GKS. The invocation of most explicit associative rules would likely be below the minimum invocation frequency, and thus they would be deleted. Therefore it would be unlikely that many rules would be established in the GKS. Thus, RBR, although present, was not significant in this task. SBR was not significant in this task either, because surface (phonological and morphological) similarity between clue words and target words was not significant.

The integration of the GKS and the AMN was as explained before: To combine the bottom-up activation of a chunk due to the AMN with the activation of a chunk due to associative rules (or external input), a *max* function was applied.

During the experimental test, clue words were presented one at a time. After the presentation of each clue word, it would be stored into the working memory. So, in effect, the partial sequence of words seen thus far was presented at each step. Thus, the activation of the target word became stronger and stronger in activation as a result of the "accumulating" input.

Due to accumulating evidence, a tentative winner (a hunch) first emerged, and then a final winner (a conviction) was generated. In the GKS, an activated chunk i was considered a hunch (a tentative winner), if $\forall j, S_i^c > S_j^c$ and $S_i^c > threshold_1$, where S_i^c indicated the strength of chunk i. A chunk i was considered a conviction (a final winner), if $\forall j, S_i^c > S_j^c$ and $S_i^c > threshold_2$. Of course, $threshold_1$ was lower than $threshold_2$, leading to partial certainty of a generated solution word. [13]

The following action rules (among many other action rules) were implemented in the ACS of CLARION for controlling this specific retrieval process in the NACS:

[11] This process was in fact the reverse of the word association test (Kent and Rosanoff 1910).

[12] For the associative memory network, the number of input units was 1600, the number of hidden units was 200, and the number of output units was 100. The learning rate was 0.3 and the momentum was 0.1.

[13] The thresholds were set at $threshold_1 = 0.036$ and $threshold_2 = 0.048$.

If goal= solving-association-task, then present the input and retrieve the most highly activated chunk and store it into working memory WM(0).

If goal= solving-association-task and $WM(0) > threshold_1$, then retrieve and report $WM(0)$ and indicate *hunch*.

If goal= solving-association-task and $WM(0) > threshold_2$, then retrieve and report $WM(0)$ and indicate *conviction*.

Similar to the previous simulation, these rules were "fixed" rules for this task, acquired presumably from a priori knowledge and task instructions given to subjects prior to experiments. When both rules were applicable, a random selection was made. In these rules, the input chunks were not involved in comparison (they were specifically excluded). The goal (which was involved in the rules) was set in the goal structure, when task instructions were given before the test began.

4.4 Simulation Results of Discovery

Recall that the human data of this task indicated that (1) the average number of clue words at which a hunch was arrived at was 10.12, with SD = 4.55; (2) the average number of clue words it took for subjects to go from a hunch to a conviction was 1.79, with SD = 0.96. Matching the human data closely, our simulation result indicated that (1) the average number of clue words at which a hunch was arrived at was 10.23, with SD = 3.07; (2) the average number of clue words it took to go from a hunch to a conviction was 1.72, with SD = 2.09. Clearly, the match was excellent.

To better understand this simulation and the interpretation of the discovery task data as embodied in this simulation, let us examine some details. First of all, from Figure 2, we see that there was a gradual accumulation of activation on the target word over time, due to successive additions of clue words. Note that this accumulation was the result of both the bottom level (the AMN) as well as the top level (the associative rules in the GKS). For example, the contribution of the top level was as shown in Figure 3, in terms of number of matching associative rules. As we can see, there was an increase in number of matching rules toward the end of the list. This led to an increasing probability of activation. However, the accumulation was also due to the bottom level, which was shown in Figure 4.

INSERT FIGURE 2 ABOUT HERE

INSERT FIGURE 3 ABOUT HERE

INSERT FIGURE 4 ABOUT HERE

During training, implicit association at the bottom level developed gradually. Figure 5 shows the gradual development of implicit associations in the AMN over time during the course of training.

INSERT FIGURE 5 ABOUT HERE

During training, explicit associative rules were formed occasionally at the top level. A sample set of rules in the GKS, extracted during training, was as follows:

Item2-Word5 $\longrightarrow$ Item2-Target
Item2-Word14 $\longrightarrow$ Item2-Target
Item3-Word3 $\longrightarrow$ Item3-Target
Item3-Word13 $\longrightarrow$ Item3-Target
Item4-Word2 $\longrightarrow$ Item4-Target
Item4-Word13 $\longrightarrow$ Item4-Target
Item5-Word13 $\longrightarrow$ Item5-Target
Item5-Word15 $\longrightarrow$ Item5-Target
Item8-Word3 $\longrightarrow$ Item8-Target
Item10-Word6 $\longrightarrow$ Item10-Target

where "item" indicated a list of words, "word" indicated a particular word at a particular position on the list, and "target" indicated the corresponding target word. As we can see, there were, in general, more rules for words toward the end of the list than toward the beginning. This was because these words toward the end were used more frequently in training, and thus were more likely to form explicit associative rules at the top level (as well as to develop stronger implicit associations with the target word in the AMN).

To further validate our model, we tested the AMN alone in simulating this task (that is, we removed

RBR altogether, through setting the balance parameter α to zero). The fit was a lot poorer, despite repeated adjustments of parameters. Comparing the AMN alone simulation with the full simulation appeared to indicate the necessity of including both types of processes (implicit and explicit) in modeling cognition in general, and in modeling this task in particular.

5 Discussions

Let us discuss the two tasks separately first. First of all, the simulation Sloman (1998) with both rule-based and similarity-based reasoning accurately captured the human data. This simulation demonstrates the importance of similarity-based reasoning (which involves both implicit and explicit processes) in human everyday reasoning. This similarity-based process is quite distinct from probabilistic reasoning as implemented in other existing cognitive architectures, such as ACT-R (see Anderson 1993 or Anderson and Lebiere 1998). Let us compare the two different approaches. ACT-R, as described in Anderson and Lebiere (1998), tries to capture all inferences in a probabilistic framework. In doing so, it lumps together all forms of weak inferential connections in a unified way. Although this approach leads to uniformity, it has shortcomings. All similarity relations between any pair of any two objects must be explicitly represented with all the associated parameters, which specify the probabilistic computation used to capture similarity-based reasoning (along with other inexact inferences). The problem is the complexity of representing all similarity pairings, which is very high in ACT-R but in contrast is not necessary in CLARION. [14]

Note that the limitations of probabilistic reasoning (Pearl 1988) in general include its neglect of many heuristics, simplifications, and rules of thumb (Tversky and Kahneman 1983, Sun 1995, Yang and Johnson-Laird 2001) useful in reducing the computational complexity of formal mathematical models. As a result, this approach suffers from higher computational complexity (see Sun 1995). Therefore, we chose not to take this approach in implementing CLARION.

We should also look into the framework of Collins and Michalski (1989), which apparently incorporated "similarity-based" reasoning through explicitly representing similarity in a complicated logical formalism. Similarity was explicitly represented as a logical operator: That is, for almost any pair of any two objects, there would be a logical relation explicitly represented, denoting their similarity. Inferences could be performed on the basis of these similarity operators using a search process. The

[14] Although partial match may be used in ACT-R to handle some similarity-based reasoning, partial match alone is not sufficient to handle the full extent of similarity-based reasoning (see Sun 1993, 1995).

complexity of this representational framework was extremely high.

Generally speaking, in terms of capturing human everyday reasoning (of which similarity-based processes and intuition are part, in our view), although logic-based models are useful, they suffer from a number of well-known shortcomings, including their restrictiveness in terms of pre-conditions, consistency, and correctness, and their inadequacy in dealing with inexactness of the real world (see, e.g., Israel 1987, Sun 1995). Their restrictiveness renders such models costly, difficult to specify, and difficult to use. These models cannot easily deal with the situations dealt with in this work, as discussed earlier in relation to the problems associated with capturing naturally similarity and strength evaluations within logics (see, e.g., section 2.1).

In a different vein, psychological work on reasoning is relevant also. Such work mostly centers around either mental logic (Rips 1994, Braine and O'Brien 1998) or mental models (Yang and Johnson-Laird 2001). These approaches, however, do not deal with the kinds of situations embodied in this work. Their focuses are elsewhere. Neither of the above two approaches have dealt with similarity-based reasoning (or intuition/insight) as captured in CLARION. They cannot easily deal with such reasoning, for essentially the same reasons as discussed earlier regarding logics.

Therefore, this line of work, combining similarity-based reasoning and rule-based reasoning (Sun 1995, Hahn and Chater 1998), offers a unique approach for capturing some essential patterns of human everyday reasoning (albeit not all patterns of human reasoning). It complements logic-based "commonsense" reasoning models prevalent in AI, which is very much centered on logic and thus also limited by logic. In addition, this approach may well be extended to case-based and/or analogical reasoning (e.g., Sun 1995a).

Second, although not a typical topic in cognitive science, intuition and insight have been documented experimentally in recently years (see, e.g., Bowers et al 1990, Schooler et al 1993, etc.). A great deal of experimental data have been accumulated on that. The explanation of this phenomenon, however, is not as clear as one would like to see, and thus computational modeling and simulation are useful.

Our simulation demonstrated the capability of CLARION in capturing and explaining this type of situation. The simulation succeeded in capturing the data without adding any new mechanisms or components—the simulation falls out of the existing mechanisms in CLARION, which include, in particular, both implicit and explicit memory.

In this regard, our model has shown that it is useful to posit the existence of these two separate

memory systems: explicit versus implicit. While explicit memory encodes rules that are all-or-nothing, implicit memory allows more gradual accumulation of information. Furthermore, the simulation has shown that the interaction between implicit and explicit memory systems, in the sense that intuition gives rise to explicit awareness and vice versa, is important to human everyday reasoning.

Overall, the two simulations were conducted based on our framework of mixed rule-based reasoning, similarity-based reasoning, and intuition (implicit associative memory), which, along with other simulations published elsewhere (e.g., Sun 1995, Sun 2002, Sun et al 2001, Sun and Zhang 2003), showed the cognitive plausibility of the CLARION architecture.

Compared with existing cognitive architectures (such as ACT-R and SOAR; Rosenbloom et al 1993, Anderson and Lebiere 1998), CLARION embodies a different set of assumptions: most notably, the separation of the two dichotomies—action-centered vs. non-action-centered knowledge, and implicit vs. explicit knowledge (Sun 2002). These alternative assumptions and structures enable CLARION to capture a variety of cognitive data (see Sun 1999, Sun et al 2001, Sun 2002 for details), including the two tasks described in the present article.

This work also points to new avenues of cognitive modeling of human everyday reasoning beyond the current psychology of reasoning (which mostly focuses on various logics and mental models). Detailed simulations help to substantiate theories and interpretations (pertaining to these reasoning experiments) and help to bring out the essential cognitive processes involved.

6 Concluding Remarks

The examination and the simulation of the two reasoning tasks, categorical inference and discovery, provide a glimpse into the inner working of human reasoning faculty, in ways that are different from existing work. This work provides an unified explanation of various reasoning patterns in a coherent way within a unified architecture. A coherent set of mechanisms were provided within the generic cognitive architecture, CLARION (Sun 2002), which succeeded in accounting for the two data sets. Thus, the simulations have validated, to some extent, the mechanisms and the postulates of CLARION concerning human reasoning.

Note that this set of studies is merely a first step in explaining and accounting for human reasoning data in a comprehensive and unified way. Our simulation so far has merely shown the promise of this approach, as well as its distinction from other approaches.

Acknowledgment

This work is supported in part by Army Research Institute contract DASW01-00-K-0012. Thanks to Robert Mathews for his collaboration on the project in general. Thanks also to Yingrui Yang and Selmer Bringsjord for their invitation to contribute to their special issue, which led to the present article.

References

J. R. Anderson, (1993). *Rules of the Mind.* Lawrence Erlbaum Associates. Hillsdale, NJ.

J. Anderson and C. Lebiere, (1998). *The Atomic Components of Thought.* Lawrence Erlbaum Associates, Mahwah, NJ.

J. Barwise and J. Etchemendy, (1995). Heterogeneous logic. In: J. Glasgow and N. Narayanan and B. Chandrasekaran (eds.), *Diagrammatic Reasoning: Cognitive and Computational Perspectives.* MIT Press, Cambridge, MA. 211-234.

K. Bowers, G. Regehr, C. Balthazard, and Parker, (1990). Intuition in the context of discovery. *Cognitive Psychology.* 22. 72-110.

M. Braine and D. O'Brien, (eds.) (1998). *Mental Logic.* Lawrence Erlbaum Associates, Mahwah, NJ.

A. Clark and A. Karmiloff-Smith, (1993). The cognizer's innards: a psychological and philosophical perspective on the development of thought. *Mind and Language.* 8 (4), 487-519.

A. Collins, (1978). Fragments of a theory of human plausible reasoning. In: D.Waltz (ed.), *Theoretical Issues in Natural Language Processing II*, 194-201. Ablex, Norword, NJ.

A. Collins and R. Michalski, (1989). The logic of plausible reasoning. *Cognitive Science*, 13(1), 1-49.

E. Davis, (1990). *Representations of Commonsense Knowledge.* Morgan Kaufman, San Mateo, CA.

U. Hahn and N. Chater, (1998). Similarity and rules: distinct? exhaustive? empirically distinguishable? *Cognition*, 65, 197-230.

D. Israel, (1987). What's wrong with non-monotonic logic? In: Ginsberg (ed.), *Readings in Non-monotonic Reasoning*, pp.53-55, Morgan Kaufman, San Mateo, CA.

G. Kent and A. Rosanoff, (1910). A study of association in insanity. *American Journal of Insanity*, 67, 37-96.

J. Metcalfe, (1986). Dynamic metacognitive monitoring during problem solving. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 12, 623-634.

M. Moscovitch and C. Umilta, (1991). Conscious and unconscious aspects of memory. In: *Perspectives on Cognitive Neuroscience*. Oxford University Press, New York.

J. Pearl, (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo, CA.

L. Rips, (1994). *The Psychology of Proof*. MIT Press, Cambridge, MA.

P. Rosenbloom, J. Laird, and A. Newell, (1993). *The SOAR Papers: Research on Integrated Intelligence*. MIT Press, Cambridge, MA.

D. Schacter, (1990). Toward a cognitive neuropsychology of awareness: implicit knowledge and anosagnosia. *Journal of Clinical and Experimental Neuropsychology*. 12 (1), 155-178.

J. Schooler, S. Ohlsson, and K. Brooks, (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, 122 (2), 166-183.

S. Sloman, (1993). Feature based induction. *Cognitive Psychology*, 25, 231-280.

S. Sloman, (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, 35, 1-33

P. Smolensky, (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11 (1), 1-74.

R. Sun, (1991). Connectionist models of rule-based reasoning. *Proceedings of the 13th Cognitive Science Conference*, pp.437-442. Lawrence Erlbaum Associates, Hillsdale, NJ.

R. Sun, (1993). An efficient feature-based connectionist inheritance scheme. *IEEE Transactions on System, Man and Cybernetics*, Vol.23, No.2. pp.512-522. 1993.

R. Sun, (1994). *Integrating Rules and Connectionism for Robust Commonsense Reasoning*. John Wiley and Sons, New York, NY. 1994.

R. Sun, (1995). Robust reasoning: Integrating rule-based and similarity-based reasoning. *Artificial Intelligence*. 75, 2. 241-296.

R. Sun, (1995a). A microfeature based approach toward metaphor interpretation. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-95)*. Montreal, Canada. pp.424-430, Morgan Kaufmann, San Francisco, CA.

R. Sun, (1999). Accounting for the computational basis of consciousness: A connectionist approach. *Consciousness and Cognition*, Vol.8, 529-565.

R. Sun, (2002). *Duality of the Mind*. Lawrence Erlbaum Associates, Mahwah, NJ.

R. Sun, E. Merrill, and T. Peterson, (2001). From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive Science*. Vol.25, No.2, 203-244.

R. Sun and T. Peterson, (1998). Autonomous learning of sequential tasks: experiments and analyses. *IEEE Transactions on Neural Networks*, Vol.9, No.6, pp.1217-1234.

R. Sun and X. Zhang, (2003). Accessibility versus action-centeredness in the representation of cognitive skills. *Proceedings of the Fifth International Conference on Cognitive Modeling*, pp.195-200. Universitats-Verlag Bamberg, Bamberg, Germany.

Y. Yang and P. Johnson-Laird, (2001). Mental models and logical reasoning problems in the GRE. *Journal of Experimental Psychology: Applied*, 7 (4), 308-316.

A. Tversky, (1977). Features of similarity. *Psychological Review*, 84(4), 327-352.

A. Tversky and D. Kahneman, (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 439-450.

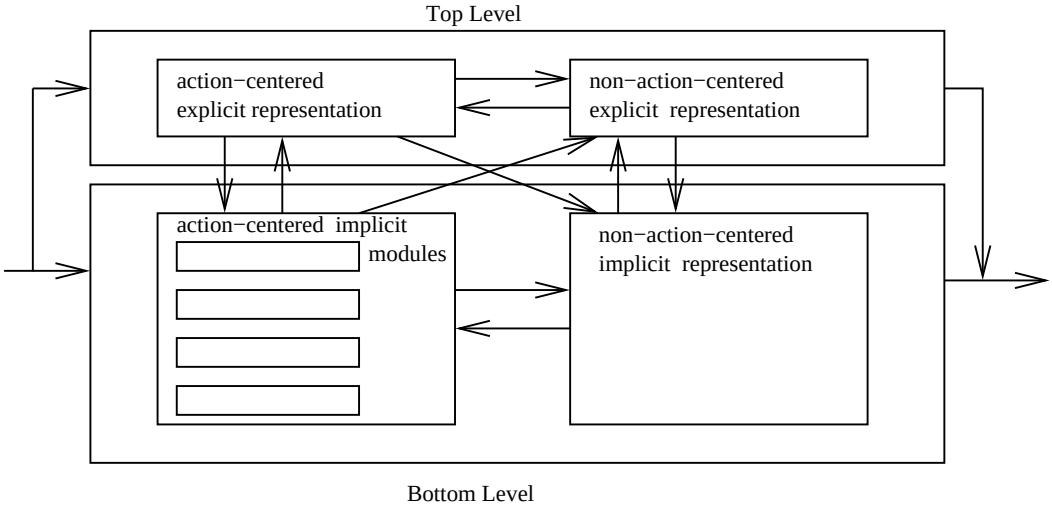

Figure 1: The CLARION architecture.

-

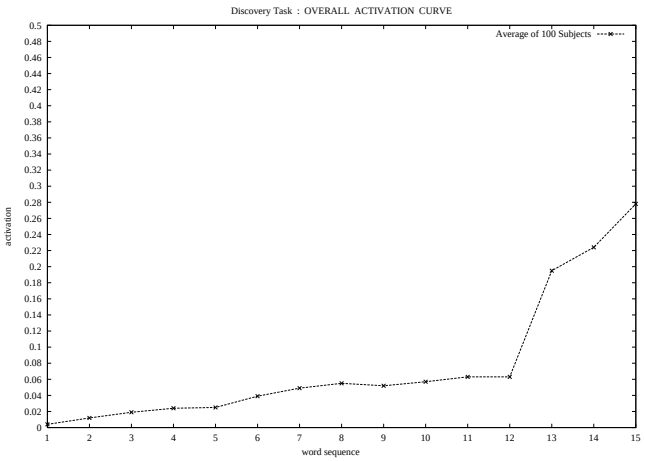

Figure 2: The accumulation of activation.

-

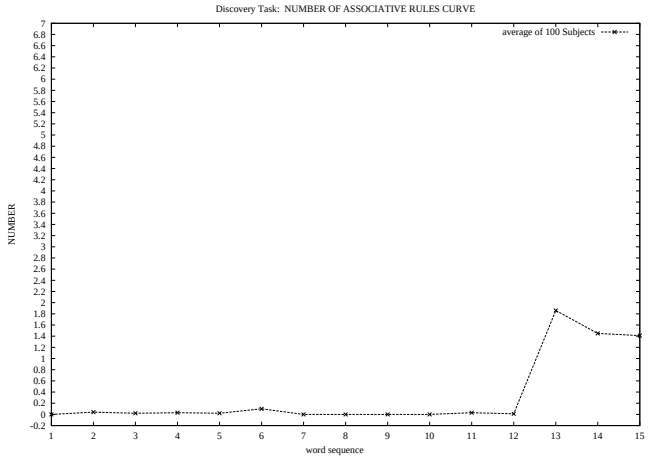

Figure 3: The number of matching rules in the GKS for each clue word.

-

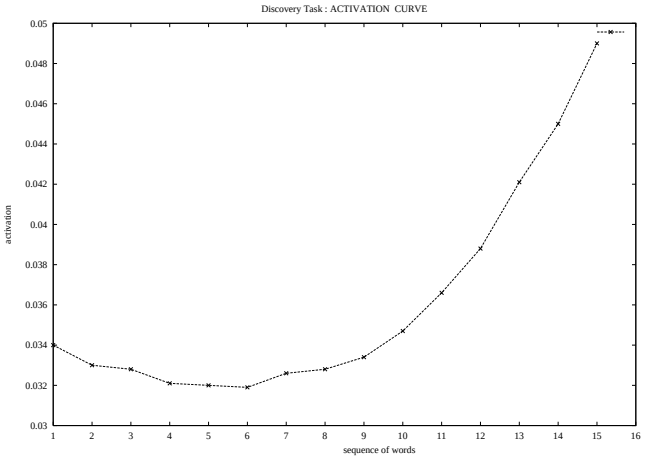

Figure 4: The accumulation of activation at the bottom level.

-

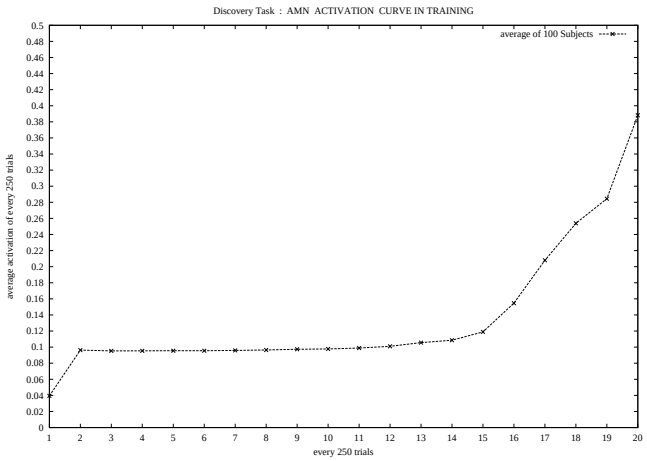

Figure 5: The training curve of the bottom level.

-