Supplementary material for "An integrative account of memory and reasoning phenomena"

Sebastien Helie (shelie@purdue.edu), Ron Sun (rsun@rpi.edu)

Appendix A: Human memory

This appendix presents some technical details on the functioning of working memory in CLARION and the modeling details of the simulations in the *Human memory* subsection. The free parameters in all the simulations were set to their default values. Before describing the simulation details, some technical details of the top-level of working memory need to be introduced.

Working memory

Top-level chunk nodes in the CLARION working memory (WM) are not organized as a network: They fill slots (there are wm_{size} slots in WM; by default, $wm_{size} = 4$). At every time step, each item in working memory has probability p of being encoded in the NACS (by default, $p = 0.1$), and one item is chosen randomly to be refreshed (which increases its base-level activation, keeping it accessible in WM).

Each WM chunk node has a base-level activation defined as (similar to ACT-R; Anderson et al., 2004):

$$b_j^c = ib_j^c + c\sum_{l=1}^{n} t_l^{-d} \qquad \text{(A1)}$$

where b_j^c is the base-level activation of chunk node j, ib_j^c is the initial base-level activation (by default, $ib_j^c = 0$), c is the amplitude (by default, $c = 2$), d is the decay rate (by default, $d = 0.5$),

and t_l is the lth use of the chunk node. This measure decays exponentially and corresponds to the odds of needing chunk j based on past experiences (Anderson, 1990).

Each new item entering WM may displace an existing chunk. The probability of being bumped is an inverse function of the (softmax) normalized base-level activations of an existing item (Sun & Helie, 2013). This is because chunks that are being used (attended to) are more likely to be needed in the future and should not be forgotten (Anderson, 1990). Specifically,

$$P(\text{bumped} = i) = \frac{1}{k\left[P(\text{chunk } i)\right]} \qquad \text{(A2)}$$

where *P(bumped = i)* is the probability that chunk i is chosen to be replaced by the new chunk, *P(chunk i)* is the probability of chunk i being needed in the future where $P(\text{chunk } i) = \frac{e^{b_i^c}}{\sum_j e^{b_j^c}}$ (A1, after softmax normalization), and k is a constant where $k = \sum_j \frac{1}{P(\text{chunk } j)}$.

Priming

The model was composed of 100 nodes, and a bipolar stimulus was randomly generated and used to train the model for 100 Epochs. One hundred and one test stimuli were generated by randomly flipping {-1, 1} the activation of between 0 and 100 nodes in the training stimulus. Each test stimulus was presented to the network and processed until convergence. The number of iterations needed for each stimulus to converge was recorded. The average number of iterations required for convergence of a random vector was computed by randomly generating 1,000 bipolar vectors. Each random vector was processed in the network until convergence, and the number of iterations was recorded. The simulation results are presented in Figure 3.

List length

The model was composed of 50 nodes. Fifty lists of stimuli were created by randomly generating between 1 and 50 bipolar stimuli (one list was generated for each list length). The model was trained with each list of stimuli for 50 Epochs. After training, each training stimulus was presented to the network, and the capacity of the network to restore the training stimulus was recorded (correct vs. incorrect). The proportion of correct recall for each list length is presented in Figure 5.

Serial position curve

The simulation of the serial position curve did not involve the attractor network in the bottom level of CLARION; only the interaction between the working memory and the NACS was required. The list lengths were: {10, 15, 20, 30, 40}. Each list length was simulated 100 times, and the proportion of recalls for each item was recorded. Because there was no time pressure, it was assumed that all the words encoded into the NACS and/or present in working memory at the end of the simulation were automatically recalled. The results are presented in Figure 6a.

The inhibition of the recency effect (Figure 6b) was simulated exactly as described above but with a list length of 15 and without recalling items from working memory (to simulate the delayed recall condition). The inhibition of the primacy effect (Figure 6c) was simulated exactly as described above with a list length of 16 words and a preloading of 3 items (as in the human experiment).

Appendix B: Deductive reasoning

This appendix presents a detailed derivation of the proofs related to the cases described in the *Deductive reasoning* subsection. In the simplest case, associative rules in CLARION can be represented using connection weights, and the top level of the NACS can be represented by a linear connectionist network (Haykin, 2009):

$$s_j^r = \sum_i s_i \times w_{ij}^r \tag{B1}$$

where s_j^r is the activation of chunk node j following the application of an associative rule, s_i is the activation of chunk node i, and w_{ij}^r is the strength of the associative rule between chunk nodes i and j (by default, $w_{ij}^r = 1/n$, where n is the number of chunk nodes in the condition of the associative rule). The application of Eq. B1 is referred to as *rule-based reasoning* (Sun, 1994).

NACS chunks also share a relationship through similarity, which enables reasoning by similarity. In CLARION, the activation of a chunk node caused by its similarity to other chunk nodes is termed *similarity-based reasoning*. Specifically,

$$s_j^s = s_{c_i \sim c_j} \times s_i \tag{B2}$$

where s_j^s is the activation of chunk node j caused by its similarity to other chunks, $s_{ci\sim cj}$ is the similarity from chunk i to chunk j, and s_i is the activation of chunk node i. The similarity measure ($s_{ci\sim cj}$) is defined through the interaction of the top and bottom levels of the NACS:

$$s_{c_i \sim c_j} = \frac{n_{c_i \cap c_j}}{f(n_{c_j})} = \frac{\sum_k w_k^{c_j} h_k(c_i, c_j)}{f\left(\sum_k w_k^{c_j}\right)} \quad \text{(B3)}$$

where w_k^{cj} is the weight of feature k in chunk j (by default, $w_k^{cj} = 1$ for all ks), $h_k(c_i, c_j) = 1$ if chunks i and j share feature k and 0 otherwise, and $f(x)$ is a slightly super linear, monotonically increasing, positive function [by default, $f(x) = x^{1.1}$]. By default, $n_{ci \cap cj}$ counts the number of features shared by chunks i and j (i.e., the feature overlap in the bottom level) and n_{cj} counts the total number of (bottom-level) features in chunk j. However, the feature weights are learned and can be varied to account for prior knowledge or context (e.g., the context emphasizes a particular feature or past experience suggests that a particular feature is more useful). Thus, similarity-based reasoning in CLARION is naturally accomplished using (1) top-down activation by chunk nodes of their corresponding bottom-level feature-based representations, (2) calculation of feature overlap between any two chunks in the bottom level (as in Eq. B3), and (3) bottom-up activation of the top-level chunk nodes (Eq. B2). This kind of similarity calculation is naturally accomplished in a multi-level cognitive architecture and represents a form of synergy between the explicit and implicit modules.

Inexact information

Let c_i and c_j be chunks in the top level of the NACS, and w_{ij}^r be a rule in the top level of the NACS linking chunks i and j. Assume that $s_i < 1$.

Derivation.

$s_j = s_i \times w_{ij}^r$

$= s_i$

In words, chunk j is partially activated, proportional to the activation of chunk i. The strength of chunk j can be used to represent the confidence in the inference.

Incomplete information

Let c_i, c_j, c_k, c_l be chunks in the top level of the NACS and w_{il}^r, w_{jl}^r, w_{kl}^r be rules in the top level of the NACS linking chunks i, j, and k with chunk l. Assume that $s_i = s_j = 1$ and that $s_k = 0$.

Derivation.

$s_l = s_i \times w_{il}^r + s_j \times w_{jl}^r + s_k \times w_{kl}^r$

$= w_{il}^r + w_{jl}^r$

$= 2/3$

In words, chunk l is partially activated, in correspondence with the proportion of its premises that are activated.

Similarity matching

Let c_i, c_j, c_k be chunks in the top level of the NACS, $s_{ci\sim cj}$ be the similarity between chunks i and j, and w_{jk}^r be a rule in the top level of the NACS linking chunks j and k. Assume that $s_i = 1$.

Derivation.

$$s_k = s_i \times s_{c_i \sim c_j} \times w_{jk}^r$$

$$= \frac{n_{c_i \cap c_j}}{f\left(n_{c_j}\right)}$$

In words, chunk k is partially activated, proportional to the similarity between chunks i and j.

Superclass to subclass inheritance

Let c_i, c_j, c_k be chunks in the top level of the NACS, the category represented by chunk i is a proper subset of the category represented by chunk j, and w_{jk}^{r} is a rule in the top level of the NACS linking chunks j and k. Assume that $s_i = 1$.

Derivation.

$$
\begin{aligned}
s_k &= s_i \times s_{c_i \sim c_j} \times w_{jk}^{r} \\
&= \frac{n_{c_i \cap c_j}}{f\left(n_{c_j}\right)} \\
&= \frac{n_{c_j}}{f\left(n_{c_j}\right)} \\
&\approx 1 \,(\text{but} < 1)
\end{aligned}
$$

In words, chunk k is activated because chunk i fully activates chunk j (up to the slight non-linearity of $f(\bullet)$, which is negligible). Chunk j has a top-level rule that transmits its activation to chunk k. This approximates the exact nature of deductive reasoning.

Subclass to superclass "inheritance"

Let c_i, c_j, c_k be chunks in the top level of the NACS, the category represented by chunk i is a proper subset of the category represented by chunk j, and w_{ik}^{r} be a rule in the top level of the NACS linking chunks i and k. Assume that $s_j = 1$.

Derivation.

$$
\begin{aligned}
s_k &= s_j \times s_{c_j \sim c_i} \times w_{ik}^{r} \\
&= \frac{n_{c_j \cap c_i}}{f\left(n_{c_i}\right)} \\
&= \frac{n_{c_j}}{f\left(n_{c_i}\right)} \\
&< 1
\end{aligned}
$$

In words, chunk *k* is partially activated, proportional to the ratio of the number of features (i.e., bottom-level nodes) of chunks *j* and *i*. Because chunk *i* represents a category that is a proper subset of the category represented by chunk *j*, chunk *i* is represented by more bottom-level features than chunk *j*. This represents the uncertainty of inductive reasoning.

Cancellation of superclass to subclass inheritance

Let c_i, c_j, c_k, c_m be chunks in the top level of the NACS, the category represented by chunk *i* is a proper subset of the category represented by chunk *j*, and w_{jk}^{r} and w_{im}^{r} are rules in the top level of the NACS linking chunks *j* and *k* and chunks *i* and *m* (respectively). Assume that $s_i = 1$.

Derivation.

$$
\begin{aligned}
s_k &= s_i \times s_{c_i \sim c_j} \times w_{jk}^{r} \\
&= \frac{n_{c_i \cap c_j}}{f\left(n_{c_j}\right)} \\
&= \frac{n_{c_j}}{f\left(n_{c_j}\right)} \\
&\approx 1 \text{ (but} < 1)
\end{aligned}
$$

while,

$$
\begin{aligned}
s_m &= s_i \times w_{im}^{r} \\
&= 1
\end{aligned}
$$

Hence, $s_m > s_k$.

In words, chunk *k* is almost fully activated, but the denominator is slightly bigger than the numerator in its derivation (because *f(•)* is super linear). In contrast, chunk *m* is fully activated, because top-level rules are exact. This shows the superiority of rule-based reasoning over similarity-based reasoning.

Cancellation of subclass to superclass "inheritance"

Let c_i, c_j, c_k, c_m be chunks in the top level of the NACS, the category represented by chunk i is a proper subset of the category represented by chunk j, and w_{ik}^{r} and w_{jm}^{r} are rules in the top level of the NACS linking chunks i and k and chunks j and m (respectively). Assume that $s_j = 1$.

Derivation.

$$
\begin{aligned}
s_k &= s_j \times s_{c_j \sim c_i} \times w_{ik}^{r} \\
&= \frac{n_{c_j \cap c_i}}{f\left(n_{c_i}\right)} \\
&= \frac{n_{c_j}}{f\left(n_{c_i}\right)} \\
&< 1
\end{aligned}
$$

$$
\begin{aligned}
s_m &= s_j \times w_{jm}^{r} \\
&= 1
\end{aligned}
$$

Hence, $s_m > s_k$.

In words, chunk k is partially activated, because chunk i has more features than chunk j (remember that chunk i represents a proper subset of chunk j). On the other hand, chunk m is fully activated, because top-level rules are exact. This restates the prominence of rule-based reasoning over similarity-based reasoning.

Mixed rules and similarities

Here, we present six subcases involving different amounts of rule-based and similarity-based reasoning.

(1) Let c_i, c_j, c_k, be chunks in the top level of the NACS, $s_{cj\sim ck}$ be the similarity between chunks j and k, and w_{ij}^{r} be a rule in the top level of the NACS between chunks i and j. Assume that $s_i = 1$.

Derivation.

$$\begin{aligned} s_k &= s_i \times w_{ij}^{r} \times s_{c_j \sim c_k} \\ &= \frac{n_{c_j \cap c_k}}{f(n_{c_k})} \end{aligned}$$

In words, chunk k is partially activated, proportional to its similarity with chunk j (because chunk j is fully activated by rule-based reasoning).

(2) Let c_i, c_j, c_k, be chunks in the top level of the NACS, and w_{ij}^{r}, w_{jk}^{r} be rules in the top level of the NACS linking chunks i and j and chunks j and k (respectively). Assume that $s_i = 1$.

Derivation.

$$\begin{aligned} s_k &= s_i \times w_{ij}^{r} \times w_{jk}^{r} \\ &= 1 \end{aligned}$$

In words, chunk k is fully activated because top-level rules are exact (i.e., rule-based reasoning is transitive).

(3) Let c_i, c_j, c_k, c_m be chunks in the top level of the NACS, $s_{ci\sim cj}$ is the similarity between chunks i and j, and w_{jk}^{r} and w_{km}^{r} are rules in the top level of the NACS linking chunks j and k and chunks k and m (respectively). Assume that $s_i = 1$.

Derivation.

$$\begin{aligned} s_m &= s_i \times s_{c_i \sim c_j} \times w_{jk}^{r} \times w_{km}^{r} \\ &= \frac{n_{c_i \cap c_j}}{f(n_{c_j})} \end{aligned}$$

In words, chunk *m* is partially activated, proportional to the similarity between chunks *i* and *j*. This is because the activation of chunk *j* is a function of its similarity to chunk *i*. However, the activation from chunk *j* to chunk *k* to chunk *m* is transmitted exactly.

(4) Let c_i, c_j, c_k, c_m be chunks in the top level of the NACS, $s_{cj\sim ck}$ is the similarity between chunks *j* and *k*, and w_{ij}^{r} and w_{km}^{r} are rules in the top level of the NACS linking chunks *i* and *j* and chunks *k* and *m* (respectively). Assume that $s_i = 1$.

Derivation.

$$\begin{aligned} s_m &= s_i \times w_{ij}^{r} \times s_{c_j \sim c_k} \times w_{km}^{r} \\ &= \frac{n_{c_j \cap c_k}}{f\left(n_{c_k}\right)} \end{aligned}$$

In words, chunk *m* is partially activated, proportional to the similarity between chunks *j* and *k*.

(5) Let c_i, c_j, c_k, c_m be chunks in the top level of the NACS, $s_{ck\sim cm}$ is the similarity between chunks *k* and *m*, and w_{ij}^{r} and w_{jk}^{r} are rules in the top level of the NACS linking chunks *i* and *j* and chunks *j* and *k* (respectively). Assume that $s_i = 1$.

Derivation.

$$\begin{aligned} s_m &= s_i \times w_{ij}^{r} \times w_{jk}^{r} \times s_{c_k \sim c_m} \\ &= \frac{n_{c_k \cap c_m}}{f\left(n_{c_m}\right)} \end{aligned}$$

In words, chunk *m* is partially activated, proportional to its similarity with chunk *k*.

(6) Let c_i, c_j, c_k, c_m be chunks in the top level of the NACS, $s_{ci\sim cj}$ and $s_{ck\sim cm}$ are similarity measures between chunks *i* and *j* and chunks *k* and *m* (respectively), and w_{jk}^{r} is a rule in the top level of the NACS between chunks *j* and *k*. Assume that $s_i = 1$.

Derivation.

$$s_m = s_i \times s_{c_i \sim c_j} \times w^r_{jk} \times s_{c_k \sim c_m}$$
$$= \frac{n_{c_i \cap c_j}}{f(n_{c_j})} \times \frac{n_{c_k \cap c_m}}{f(n_{c_m})}$$

This case is a little more complex to interpret. Chunk *m* is partially activated, proportional to its similarity to chunk *k*. However, unlike in the previous cases, chunk *k* is not fully activated: the activation of chunk *k* is a function of the similarity between chunks *i* and *j*. Hence, all else being equal, the activation of chunk *m* is smaller here than in the preceding subcase (5).

Appendix C: The role of functional attributes in inductive reasoning

In CLARION, induction is accounted for by similarity-based reasoning. The reader is referred to Eqs. B2-B3 in Appendix B above for a formal treatment of similarity-based reasoning in CLARION. Only one numerical parameter was varied to derive the following role of functional attributes (i.e., w_k^{cj}, the weight of feature k in chunk j). The relative weights of features in CLARION can be emphasized by the context (e.g., compare concepts x and y according to their color) or learned through past experience (e.g., learning what feature is useful in a specific task using reinforcement learning).

Let chunk i represent 'chicken', chunk t represent 'tiger', and chunk j represent 'hawk'. If all the features are weighed equally, formalization using Eqs. B2 and B3 yields:

$$s_i \times \frac{n_{c_i \cap c_j}}{f(n_{c_j})} > s_t \times \frac{n_{c_t \cap c_j}}{f(n_{c_j})} \Rightarrow s_i \times \frac{\sum_k w_k^{c_j} h_k(c_i, c_j)}{f\left(\sum_k w_k^{c_j}\right)} > s_t \times \frac{\sum_k w_k^{c_j} h_k(c_t, c_j)}{f\left(\sum_k w_k^{c_j}\right)}$$

where $h_k(c_i, c_j) = 1$ if chunks i and j share feature k and 0 otherwise, and w_k^{cj} is the weight of feature k in chunk j. The denominators are the same so they can be dropped. Also, let feature k = 0 represent the functional attribute of feeding habit:

$$s_i \times \left[w_0^{c_j} \times h_0(c_i, c_j) + \sum_{k>0} w_k^{c_j} h_k(c_i, c_j) \right] > s_t \times \left[w_0^{c_j} \times h_0(c_t, c_j) + \sum_{k>0} w_k^{c_j} h_k(c_t, c_j) \right]$$

Because $k = 0$ represents feeding habits, $h_0(c_i, c_j) = 0$ and $h_0(c_t, c_j) = 1$.

$$s_i \times \sum_{k>0} w_k^{c_j} h_k(c_i, c_j) > s_t \times \left[w_0^{c_j} + \sum_{k>0} w_k^{c_j} h_k(c_t, c_j) \right]$$

The above expression represents a regular case of similarity effect in induction. What is the condition that would reverse this inequality and create an exception to similarity?

$$s_i \times \sum_{k>0} w_k^{c_j} \; h_k\left(c_i, c_j\right) < s_t \times \left[w_0^{c_j} \; + \sum_{k>0} w_k^{c_j} \; h_k\left(c_t, c_j\right) \right]$$

$$\Rightarrow \; w_0^{c_j} \; > \frac{s_i}{s_t} \times \sum_{k>0} w_k^{c_j} \; h_k\left(c_i, c_j\right) - \sum_{k>0} w_k^{c_j} \; h_k\left(c_t, c_j\right)$$

Hence, an exception to the similarity effect can be observed when the weight of a feature (here feeding habit) is larger than the difference between the weighted feature overlap (excluding feeding habit) of the two induction cases (where the feature overlap of the normal case is modulated by chunk activation). This situation is mathematically described by the derived inequality above. In this way, CLARION accounts for exceptions to similarity in induction when the context sufficiently emphasizes a particular exception feature.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036-1060.

Haykin, S. (2009). *Neural Networks and Learning Machines. 3rd Edition*. Upper Saddle River, NJ: Prentice-Hall.

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

Sun, R., & Hélie, S. (2013). Psychologically realistic cognitive agents: Taking human cognition seriously. *Journal of Experimental & Theoretical Artificial Intelligence*, *25*, 65-92.