

## Neural Networks Letter

# Cogent confabulation

Robert Hecht-Nielsen

Computational Neurobiology, Institute for Neural Computation, Electrical and Computer Engineering Department,  
University of California, San Diego, La Jolla, CA 92093-0407, USA

Received 1 September 2004; accepted 3 November 2004

### Abstract

A new model of vertebrate cognition is introduced: maximization of *cogency*  $p(\alpha\beta\gamma\delta|\varepsilon)$ . This model is shown to be a direct generalization of Aristotelian logic, and to be rigorously related to a calculable quantity. A key aspect of this model is that in Aristotelian logic information environments it functions logically. However, in non-Aristotelian environments, instead of finding the conclusion with the highest probability of being true (a popular past model of cognition); this model instead functions in the manner of the ‘duck test,’ by finding that conclusion which is most supportive of the truth of the assumed facts.

© 2004 Elsevier Ltd. All rights reserved.

*Keywords:* Vertebrate cognition; Symbolic representations; Attractor neural network; Cogency; Confabulation; Duck test; Inductive logic; AI

### 1. Introduction

An appealing model of cognition (Bender, 1996; Nilsson, 1998; Pearl, 2000) is to generalize Aristotelian implication  $\alpha\beta\gamma\delta \Rightarrow \varepsilon$  by finding that symbol  $\varepsilon$  which maximizes a posteriori probability  $p(\varepsilon|\alpha\beta\gamma\delta)$  (for concreteness, four *assumed fact* symbols  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$ , and a conclusion symbol  $\varepsilon$ , each drawn from its own separate lexicon, with juxtaposition indicating Boolean AND, will be used in the discussion of this letter; the generalization to arbitrary situations is obvious). However, as discussed in Section 4 below, this model of cognition is not correct. This letter introduces a new model of vertebrate cognition: maximization of *cogency*  $p(\alpha\beta\gamma\delta|\varepsilon)$  and considers some related mathematical quantities.

*Some terminology.* Assuming that the combined assumed facts  $\alpha\beta\gamma\delta$  are true, the set of all symbols  $\lambda$  (in the *answer lexicon* from which conclusions are being sought) with  $p(\alpha\beta\gamma\delta|\lambda) > 0$  is called the *expectation*; the elements of which, in descending order of their cogencies, are termed *answers*.

### 2. Cogency and confabulation

Assume that  $\alpha\beta\gamma\delta \Rightarrow \varepsilon$  exclusively in the answer lexicon. Then  $p(\alpha\beta\gamma\delta\varepsilon) > 0$  and  $p(\alpha\beta\gamma\delta\lambda) = 0$  for all such other answer lexicon symbols  $\lambda$ . Thus,  $p(\alpha\beta\gamma\delta|\varepsilon) = p(\alpha\beta\gamma\delta\varepsilon)/p(\varepsilon) > 0$  and  $p(\alpha\beta\gamma\delta|\lambda) = p(\alpha\beta\gamma\delta\lambda)/p(\lambda) = 0$  for all other symbols  $\lambda$ .

This establishes:

**Theorem 1.** *If  $\alpha\beta\gamma\delta \Rightarrow \varepsilon$  exclusively, then maximization of cogency produces one and only one answer:  $\varepsilon$ .*

Thus, surprisingly, in an Aristotelian logic information environment, maximizing cogency will produce logical answers. But what about more general environments? Conceptually, cogency maximization works like the *duck test*: if a duck-sized creature quacks like a duck, walks like a duck, swims like a duck, and flies like a duck (assumed facts  $\alpha\beta\gamma\delta$ ), then we accept it as a duck (because duck,  $\varepsilon$ , is the symbol that, when it is seen, most strongly supports the probability of these assumed facts being true; i.e.  $\varepsilon$  maximizes  $p(\alpha\beta\gamma\delta|\varepsilon)$ ). There is no logical guarantee that this creature is a duck; but maximization of cogency makes the decision that it is and moves on.

Of course, cogency  $p(\alpha\beta\gamma\delta|\varepsilon)$  is a conceptual, notional quantity and can only be calculated in trivial situations.

E-mail address: [rh-n@ucsd.edu](mailto:rh-n@ucsd.edu)

Consider the possibility of using *confabulation* (maximization of the product  $p(\alpha|\varepsilon) \cdot p(\beta|\varepsilon) \cdot p(\gamma|\varepsilon) \cdot p(\delta|\varepsilon)$  or, equivalently, the sum of the logarithms of these probabilities) as a surrogate for maximizing cogency (it is assumed that all required pairwise conditional probabilities  $p(\psi|\lambda)$  between symbols  $\psi$  and  $\lambda$  are known. This assumption is termed *exhaustive knowledge*). Each meaningful non-zero  $p(\psi|\lambda)$  is termed an individual *item of knowledge*.

An exact mathematical relationship between the confabulation product and cogency is now derived. Applying the probabilistic chain rule identity  $p(abcde) = p(a|bcde) \cdot p(b|cde) \cdot p(c|de) \cdot p(d|e) \cdot p(e)$  to cogency, and using the fact that the AND operation commutes, we can write the quantity  $p(\alpha\beta\gamma\delta|\varepsilon)$  in all four of the following ways:

$$\begin{aligned} p(\alpha\beta\gamma\delta|\varepsilon) &= p(\alpha\beta\gamma\delta\varepsilon)/p(\varepsilon) \\ &= p(\alpha|\beta\gamma\delta\varepsilon) \cdot p(\beta|\gamma\delta\varepsilon) \cdot p(\gamma|\delta\varepsilon) \cdot p(\delta|\varepsilon) \end{aligned}$$

$$\begin{aligned} p(\alpha\beta\gamma\delta|\varepsilon) &= p(\beta\gamma\delta\alpha\varepsilon)/p(\varepsilon) \\ &= p(\beta|\gamma\delta\alpha\varepsilon) \cdot p(\gamma|\delta\alpha\varepsilon) \cdot p(\delta|\alpha\varepsilon) \cdot p(\alpha|\varepsilon) \end{aligned}$$

$$\begin{aligned} p(\alpha\beta\gamma\delta|\varepsilon) &= p(\gamma\delta\alpha\beta\varepsilon)/p(\varepsilon) \\ &= p(\gamma|\delta\alpha\beta\varepsilon) \cdot p(\delta|\alpha\beta\varepsilon) \cdot p(\alpha|\beta\varepsilon) \cdot p(\beta|\varepsilon) \end{aligned}$$

$$\begin{aligned} p(\alpha\beta\gamma\delta|\varepsilon) &= p(\delta\alpha\beta\gamma\varepsilon)/p(\varepsilon) \\ &= p(\delta|\alpha\beta\gamma\varepsilon) \cdot p(\alpha|\beta\gamma\varepsilon) \cdot p(\beta|\gamma\varepsilon) \cdot p(\gamma|\varepsilon) \end{aligned}$$

Multiplying these equations together gives:

$$\begin{aligned} [p(\alpha\beta\gamma\delta|\varepsilon)]^4 &= [p(\alpha|\beta\gamma\delta\varepsilon) \cdot p(\beta|\gamma\delta\varepsilon) \cdot p(\gamma|\delta\varepsilon)] \cdot [p(\beta|\gamma\delta\alpha\varepsilon) \cdot \\ &\quad p(\gamma|\delta\alpha\varepsilon) \cdot p(\delta|\alpha\varepsilon)] \cdot [p(\gamma|\delta\alpha\beta\varepsilon) \cdot p(\delta|\alpha\beta\varepsilon) \cdot \\ &\quad p(\alpha|\beta\varepsilon)] \cdot [p(\delta|\alpha\beta\gamma\varepsilon) \cdot p(\alpha|\beta\gamma\varepsilon) \cdot p(\beta|\gamma\varepsilon)] \cdot \\ &\quad [p(\alpha|\varepsilon) \cdot p(\beta|\varepsilon) \cdot p(\gamma|\varepsilon) \cdot p(\delta|\varepsilon)]. \end{aligned}$$

Applying Bayes' law to the conditional probabilities in the first four parentheses yields:

$$\begin{aligned} [p(\alpha\beta\gamma\delta|\varepsilon)]^4 &= [p(\alpha\beta\gamma\delta\varepsilon)/p(\beta\gamma\delta\varepsilon) \cdot p(\beta\gamma\delta\varepsilon)/p(\gamma\delta\varepsilon) \cdot \\ &\quad p(\gamma\delta\varepsilon)/p(\delta\varepsilon)] \cdot [p(\beta\gamma\delta\alpha\varepsilon)/p(\gamma\delta\alpha\varepsilon) \cdot \\ &\quad p(\gamma\delta\alpha\varepsilon)/p(\delta\alpha\varepsilon) \cdot p(\delta\alpha\varepsilon)/p(\alpha\varepsilon)] \cdot \\ &\quad [p(\gamma\delta\alpha\beta\varepsilon)/p(\delta\alpha\beta\varepsilon) \cdot p(\delta\alpha\beta\varepsilon)/p(\alpha\beta\varepsilon) \cdot \\ &\quad p(\alpha\beta\varepsilon)/p(\beta\varepsilon)] \cdot [p(\delta\alpha\beta\gamma\varepsilon)/p(\alpha\beta\gamma\varepsilon) \cdot \\ &\quad p(\alpha\beta\gamma\varepsilon)/p(\beta\gamma\varepsilon) \cdot p(\beta\gamma\varepsilon)/p(\gamma\varepsilon)] \cdot [p(\alpha|\varepsilon) \cdot \\ &\quad p(\beta|\varepsilon) \cdot p(\gamma|\varepsilon) \cdot p(\delta|\varepsilon)]. \end{aligned}$$

If any of the probabilities within the first four bracketed quantities on the right side of this equation are zero, but the fifth bracketed quantity is not zero, then this is said to be an *exceptional* case. Noting that the first probability in each of the first four parentheses equals  $p(\alpha\beta\gamma\delta\varepsilon)$  and rearranging and simplifying yields:

**Theorem 2.** *Given non-exceptional assumed facts  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$ , and expectation element  $\varepsilon$ , then the following exact relationship holds between cogency  $p(\alpha\beta\gamma\delta|\varepsilon)$  and the confabulation product  $p(\alpha|\varepsilon) \cdot p(\beta|\varepsilon) \cdot p(\gamma|\varepsilon) \cdot p(\delta|\varepsilon)$ :*

$$\begin{aligned} [p(\alpha\beta\gamma\delta|\varepsilon)]^4 &= [p(\alpha\beta\gamma\delta\varepsilon)/p(\alpha\varepsilon)] \cdot [p(\alpha\beta\gamma\delta\varepsilon)/p(\beta\varepsilon)] \cdot \\ &\quad [p(\alpha\beta\gamma\delta\varepsilon)/p(\gamma\varepsilon)] \cdot [p(\alpha\beta\gamma\delta\varepsilon)/p(\delta\varepsilon)] \cdot \\ &\quad [p(\alpha|\varepsilon) \cdot p(\beta|\varepsilon) \cdot p(\gamma|\varepsilon) \cdot p(\delta|\varepsilon)]. \end{aligned}$$

To see a key implication of Theorem 2, consider the following concrete case: five distinct, but identical, lexicons. Each lexicon possesses 10,000 symbols; each representing exactly one of the 10,000 most common words in a huge reference corpus of uncapitalized proper English text (novels, encyclopedias, news stories, etc.). Let assumed fact symbols  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$  be drawn from lexicons 1, 2, 3, and 4, respectively; representing a contiguous sequence of four words of text. Then let us consider which symbols  $\varepsilon$ , if any, from lexicon 5 would make a suitable completion to this phrase  $\alpha\beta\gamma\delta$ . These  $\varepsilon$ 's will be the symbols with the largest cogencies  $p(\alpha\beta\gamma\delta|\varepsilon)$ . To make the example even more concrete, let the assumed fact phrase be:  $\alpha\beta\gamma\delta =$  the train was going and consider one possible expectation symbol  $\varepsilon$ , representing the word south. Then

$$\begin{aligned} p(\alpha\beta\gamma\delta\varepsilon)/p(\alpha\varepsilon) &= p(\text{the train was going south})/ \\ &\quad p(\text{the } \_ \_ \_ \text{ south}) \\ p(\alpha\beta\gamma\delta\varepsilon)/p(\beta\varepsilon) &= p(\text{the train was going south})/ \\ &\quad p(\_ \text{ train } \_ \_ \text{ south}) \\ p(\alpha\beta\gamma\delta\varepsilon)/p(\gamma\varepsilon) &= p(\text{the train was going south})/ \\ &\quad p(\_ \_ \text{ was } \_ \text{ south}) \end{aligned}$$

and

$$\begin{aligned} p(\alpha\beta\gamma\delta\varepsilon)/p(\delta\varepsilon) &= p(\text{the train was going south})/ \\ &\quad p(\_ \_ \_ \text{ going south}) \end{aligned}$$

where an underline indicates a word position that is not being considered in calculating the probability.

Note that if  $\varepsilon$  (south) were replaced by any other expectation element (north, east, west, fast, slow, etc.) these ratios would probably change very little. Thus, in non-exceptional cases, these first four terms might function approximately as a positive constant independent of  $\varepsilon$ ; making the product  $p(\alpha|\varepsilon) \cdot p(\beta|\varepsilon) \cdot p(\gamma|\varepsilon) \cdot p(\delta|\varepsilon)$  and the fourth power of cogency approximately proportional. Under these circumstances, confabulation and maximizing cogency will give the same answers. Theorem 2 is postulated to be the 'fundamental theorem' of vertebrate cognition.

While the above argument may be correct in the specific case of English phrase completion considered, how can we be sure that the first four terms of the fundamental theorem will, in general, be approximately constant for all expectation elements  $\varepsilon$ ? In general, we cannot. Besides the problem of exceptions (which, it turns out, can be handled

by explicitly learning all of them), badly designed lexicons or ill-mannered information environments can almost certainly cause these first four terms to not be approximately constant for all expectation elements.

Biological systems find ways of exploiting a variety of scientific, technological, and mathematical principles; but to do so, they must often improvise (by evolution) specific designs that conform to the requirements and limitations of the principle. Cellular biochemistry developed in the ocean and so we must carry the ocean around with us in order to use these innovations. Cognition is presumably like this. With the proper lexicons and knowledge; developed in the proper sequence via exposure to the proper information environments (which are all things that, ultimately, genetics, and therefore evolution, can control); the fundamental theorem of cognition can be exploited and confabulation can be *cogent*. Without these restrictions, confabulation probably does not work.

### 3. Confabulation examples

Here are some examples of confabulation applied to phrase completion. As in the discussion of Theorem 2, a sequence of four words, each from its own lexicon, are given as assumed facts  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$  (in some examples only one  $\delta$ , two  $\gamma\delta$ , or three  $\beta\gamma\delta$ , assumed facts are used). Confabulation is used to select the symbol  $\varepsilon$  of the next word after the assumed fact phrase. Each lexicon has 10,000 symbols; representing the 10,000 most commonly encountered words in a  $1.4 \times 10^9$  word proper English training corpus composed of books, news stories, encyclopedias, etc. For simplicity, capital letters are not used. The pairwise conditional probabilities were obtained by marching a five-contiguous-word window one word at a time from the beginning word of the training corpus to the end. Each time a contiguous substring of two to five words without internal punctuation appeared at the right-hand end of the window, counts were gathered and stored for that substring in the corresponding selections of four  $10,000 \times 10,000$  matrices (one for each of the first four symbol lexicons paired with the fifth).

After the march through the training corpus, the probability  $p(\psi|\lambda)$  between symbols  $\psi$  and  $\lambda$  was approximated by  $c(\psi,\lambda)/c(\lambda)$ , where  $c(\psi,\lambda)$  is the count of the number of times the word represented by symbol  $\psi$  (belonging to one of the first four lexicons) and the word represented by symbol  $\lambda$  (belonging to the fifth lexicon) appeared together; and  $c(\lambda)$  is the total number of times symbol  $\lambda$  appeared in lexicon five with any symbol of the  $\psi$  lexicon during the march ( $c(\lambda)$  is equal to the sum of the  $c(\phi,\lambda)$  across all symbols  $\phi$  belonging to the  $\psi$  lexicon). Only the pairwise conditional probabilities between symbols in the first four lexicons, conditioned on symbols of the fifth lexicon, were computed because the phrase completion confabulations were always carried out with the fifth lexicon

as the answer lexicon. Symbol pair counts below 3 were thrown out (set to zero) as accidental or meaningless; as were calculated  $p(\psi|\lambda)$  values below 0.001. This knowledge acquisition process yielded 5,251,335 meaningful items of knowledge. This knowledge acquisition process was carried out on a desktop computer in a few hours.

To test the system, a highly literate, but non-technical person was asked to create a number of test word sequences to be completed. Confabulation was then applied to these assumed facts with the results shown below. The symbols determined by confabulation to be expectation elements are shown in decreasing order of the confabulation product value. When an expectation had seven or more answers, the words corresponding to the top six are shown in parentheses, followed by the total number of symbols in the expectation; if six or fewer answers were found, square brackets are used, and all of the corresponding words are shown:

- she could determine (whether, exactly, if, why, how, precisely) 8
- if it was not (immediately, clear, enough, true, properly, stupid) >999
- earthquake activity was [centered]
- for lack of a (unified, blockbuster, comprehensive, definitive, coordinated, protein) 111
- a lack of (urgency, oxygen, understanding, confidence, communication, enthusiasm) 407
- regardless of expected [outcome, length]
- cars drove down a (lane, freeway, highway, dirt, taxi, tying) 9
- driving west on interstate [highway, freeway]
- snow fell in (freezing, montana, portions, northwestern, northeastern) 11
- the facts point to [ ]
- threats of terrorist [attacks, retaliation, strikes, violence]
- the machine (tools, tool, guns, gun, operator, shop) 33
- children can learn [lessons, math, english]
- students can learn [lessons, math, english]
- college students can learn [math]
- knowledge of historical [facts, subjects, styles]
- questions that cannot be (answered, solved, resolved, avoided, addressed, yes) 9
- benefits from additional (cost-cutting, taxable, protections, taxes, acquisitions, payroll) 11
- limitations [expired, expires, imposed]

- her responsibility for taking [sole, matters]
- her responsibility for making [errors, matters, sure, references, choices, lethal]
- his responsibility for making (mistakes, matters, bombs, references, decisions, sure) 11
- his responsibility for taking [actions, sole, matters, decisions]
- mechanical failure [caused]
- crowded (commuter, marketplace, subway, courtroom, skies, sidewalk) 62
- they crowded (onto, lobby, shopping, shelters, around, into) 22
- beaches are covered with [pools]
- there were many (indications, surprises, instances, casualties, signs, exceptions) >999
- are easy to (install, dismiss, detect, locate, accumulate, criticize) 159
- microsoft makes software for [apple's, desktop, hardware]
- the green car turned [yellow].

Notice the automatic and instantaneous emergence of 'grammar' and 'semantics.'

#### 4. Discussion

In the experiments above, many of the completed phrases never appeared anywhere in the training corpus. This strong ability to correctly generalize to cases which are novel in detail, but which involve familiar elements and include no unlearned exceptions, is a favorable characteristic that confabulation seems to possess. In addition, tens of nonsense phrases (e.g.  $\alpha\beta\gamma\delta = \text{tune card fly bold}$ ) were tested as assumed facts for phrase completion and, in every case, confabulation returned no answers.

Some of the phrase completion examples presented above are perfectly valid sets of assumed facts (e.g. the facts point to), which should have at least some reasonable completions. Yet confabulation returns no answers. This is an indication that the knowledge used is not completely exhaustive. Thus, a collection of knowledge that falls somewhat short of being exhaustive seems to translate into a tendency to make errors of omission, not commission; a generally favorable attribute.

Synaptic learning of the  $p(\psi|\lambda)$  probabilities (Hecht-Nielsen, 2004) and the high-speed parallel competitive 'winner take all' implementation of confabulation by neuronal attractor networks (Amit, 1989; Anderson, Silverstein, Ritz, & Jones, 1977; Sommer & Palm, 1999) seem biologically plausible.

Now, as quickly as you can, select a next word for each of the following phrases:

- company rules forbid taking
- mickey and minnie were
- capitol hill observers are
- paper is made from
- riding the carousel was

The idea of finding that conclusion  $\varepsilon$  which has the highest probability of being true,  $p(\varepsilon|\alpha\beta\gamma\delta)$ , given the assumed facts,  $\alpha\beta\gamma\delta$ , has, for decades, been an attractive model of cognition. This attractiveness is seemingly bolstered by the (average error rate) optimality of a posteriori probability in pattern classification. A great deal of study has gone into this model (Bender, 1996; Nilsson, 1998; Pearl, 2000). However, this is actually an awful model of cognition. Consider the following numerical calculations.

Given a set of assumed facts  $\alpha\beta\gamma\delta$ , let  $\lambda$  be a conclusion with a priori probability  $p(\lambda)=0.01$  and  $\varepsilon$  an alternative conclusion with  $p(\varepsilon)=0.0001$ . Also assume that  $p(\alpha\beta\gamma\delta|\lambda)=0.01$  and  $p(\alpha\beta\gamma\delta|\varepsilon)=0.2$ . Applying Bayes' law twice to a posteriori probability yields:

$$p(\psi|\alpha\beta\gamma\delta) = p(\alpha\beta\gamma\delta|\psi) \cdot [p(\psi)/p(\alpha\beta\gamma\delta)].$$

Thus,  $p(\lambda|\alpha\beta\gamma\delta)=5 \cdot p(\varepsilon|\alpha\beta\gamma\delta)$  and so the policy of maximizing a posteriori probability will overwhelmingly choose  $\lambda$  over  $\varepsilon$ ; even though  $p(\alpha\beta\gamma\delta|\varepsilon)=20 \cdot p(\alpha\beta\gamma\delta|\lambda)$ . Thus, it should be possible to discern whether maximum a posteriori probability is a good model of cognition by seeing how strongly a priori probability enters into cognitive decision-making. In performing the above phrase completions, you have produced relevant (although perhaps not statistically significant!) experimental data that bears on this question.

In an informal poll, typical answers for the completions were: naps, happy, wondering, wood, and fun. However, in each example, the word *the* is both a viable answer and, by far, the most frequent word in English; so, if maximization of a posteriori probability were a correct theory of cognition, you would surely have selected it, overwhelmingly (as the above calculation illustrates), in every case. Instead, as with the computer confabulation experiments presented in Section 3, you probably selected words with much higher cogencies than *the*.

Cogency,  $p(\alpha\beta\gamma\delta|\varepsilon)$ , the cognitive analog of class probability in pattern classification theory, and likelihood in probability and statistics, has always been there. Waiting. Perhaps its time has finally arrived.

#### Acknowledgements

Thanks to Adrian T. Fan, Kate Mark, Robert W. Means, and Syrus C. Nemat-Nasser for implementing

the confabulation experiments and to Fair Isaac Corporation for long-term support of this research.

## References

- Amit, D. (1989). *Modeling brain function: The world of attractor networks*. Cambridge, UK: Cambridge University Press.
- Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, 84, 413–451.
- Bender, E. A. (1996). *Mathematical methods in artificial intelligence*. Los Alamitos, CA: IEEE Computer Society Press.
- Hecht-Nielsen, R. (2004). A theory of cerebral cortex. *Institute for Neural Computation, University of California, San Diego, Technical Report #0404*. Available: <http://inc2.ucsd.edu>.
- Nilsson, N. J. (1998). *Artificial intelligence: A new synthesis*. San Francisco: Morgan Freeman Publishers.
- Pearl, J. (2000). *Causality*. Cambridge, UK: Cambridge University Press.
- Sommer, F. T., & Palm, G. (1999). Improved bidirectional retrieval of sparse patterns stored by Hebbian learning. *Neural Networks*, 12, 281–297.