Beyond Copycat: Toward a Self-Watching Architecture for High-Level Perception and Analogy-Making

James B. Marshall and **Douglas R. Hofstadter**
Center for Research on Concepts & Cognition
Department of Computer Science
Indiana University
510 North Fess Street
Bloomington, IN 47408
{jim,dughof}@cogsci.indiana.edu

This report summarizes recent continuing work on the Copycat project, a stochastic computer model of fluid concepts, high-level perception, and analogy-making developed by Hofstadter and Mitchell (Mitchell, 1993). Copycat perceives analogies between short strings of letters, which can be thought of as representing abstract situations in an idealized microworld. An example of an analogy problem taken from this microworld might be "If **abc** changes to **abd**, how does **srqp** change in an analogous way?" An interesting feature of such problems is that there is no single "right" answer; rather, a range of answers is always possible for each problem. For the previous example, some possible answers might be **srqo**, **trqp**, **srqd**, **drqp**, or even **abd**. Of course, some answers are consistently judged by people to be better than others, for most analogy problems. Furthermore, for some problems, the answers judged to be the "best" are not at all the most "obvious" ones.

Copycat's nondeterministic, stochastic processing mechanisms allow it to find a number of different answers to a given analogy problem, and in its current stage of development, the program is quite good at reproducing the range and frequencies of answers given by people to certain problems, where an answer's frequency corresponds to its "obviousness". The model also incorporates a simple numerical measure of answer quality which agrees well with the relative judgments of answer quality given by people for certain problems.

Unfortunately, such a stark, numerical measure is extremely crude, and reflects a fundamental weakness of the current model: its almost complete lack of any in-depth understanding of the answers it finds. Copycat is unable to explain *why* it considers particular answers to be good or bad. The reason is that Copycat's processing mechanisms focus almost exclusively on perceiving patterns and relationships in the perceptual *data* (the letter strings), while ignoring patterns that occur in its own *processing* when solving an analogy problem. Thus, although it may discover an insightful answer for some problem, it lacks any internal representation or knowledge of the underlying process that led it to discover that answer—knowledge that could provide a basis for explaining the answer's relative strengths or weaknesses, thereby permitting a much richer assessment of its quality. Copycat's lack of any such "self-watching" ability stands in marked contrast to people, who are typically able to give an account of why they consider one answer to be better or worse than another for a particular analogy problem. An interesting related phenomenon, dubbed the *self-explanation effect*, has been studied recently in the context of students learning to solve physics problems from worked-out examples (Chi *et al.*, 1989).

Current work on Copycat is focused on developing mechanisms to allow the program to perceive and remember important processing events that occur as it works on an analogy problem—such as the recognition of key concepts or similarities that arise when the problem is viewed in a particular way—and to create explicit representations of these events. These representations, called *themes*, provide an explicit temporal trace of the program's "train of thought" as it searches for an answer, and can then be stored in memory along with an answer when one is found. In some ways, this approach is similar in flavor to work on derivational analogy (Carbonell, 1986). However, the focus here is not on improving system performance by learning to make *better* analogies, but rather on being able to explain *why* one analogy is judged to be more compelling than another.

Enriching the model's understanding of its answers by incorporating higher-order thematic information gleaned from self-watching should enable Copycat to perceive abstract similarities and differences among the analogies it makes. It should be able to apply the same processing mechanisms that it now uses to perceive relationships in its perceptual input to the more abstract "meta-level" task of perceiving relationships among its answers stored in memory, comparing and contrasting them in a way not currently possible. In short, it should eventually be able to make analogies between analogies. Endowing Copycat with a sophisticated self-watching capability forms the central theme of present efforts to extend and refine the model, and is a logical next step along the road to understanding and capturing the full richness of high-level perception and analogy-making in a compu-tational framework.

References

Carbonell, J. (1986). Derivational analogy. In R. Michalski, J. Carbonell & T. Mitchell, (Eds.), *Machine Learning: An Artificial Intelligence Approach, Volume II* (pp. 371-392). Morgan Kaufmann.

Chi, M., Bassok, M., Lewis, M., Reimann, P. & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145-182.

Mitchell, M. (1993). *Analogy-Making as Perception.* Cambridge, MA: MIT Press/Bradford Books.