The Ineradicable Eliza Effect and Its Dangers

Salient Sentence Strings: Fluid Concepts and Creative Analogies (p.155-p.168)

1. In the meantime, how- ever, a great deal of uncritical publicity was being given to a number of AI programs that gave the appearance of creating very complex real-world analogies or making sophisticated scientific discoveries, rivaling in insight such pioneers as Galileo, Kepler, and Ohm. Such favorable publicity could not help but make our achievements in micro-domains look rather microscopic, by comparison. And yet, we felt that any such conclusion about our work would be superficial and unwarranted.

2. There is an insidious problem in writing about such a computer achievement, however. When someone writes, or reads "the program makes analogy between heat flow through a metal bar and water flow through a pipe", there is a tacit acceptance that the computer is really dealing with the idea of heat flow, the idea of water flow, the concepts of heat, water, metal bar, pipe, and so on. Otherwise, what would it mean to say that it had "made an analogy"?

3. This type of illusion is generally known as the "Eliza Effect", which could be defined as the susceptibility of people to read far more understanding than is warranted into strings of symbols – specially words – strung together by computers. A trivial example of this effect might be someone thinking that an automatic teller machine really was grateful for receiving a deposit slip, simply because it printed out "THANK YOU" on its little screen.

4. Please understand that what I am saying is not meant as criticism of the developer of SEM, or even of Waldrop. It is meant much more as a critique of the whole mentality swirling around the complex intellectual endeavor called "AI" -a surprisingly unguarded mentality in which anthropomorphic characterizations of what computers do are accepted far too easily, both outside and within the field.

5. The program has been informed of nothing – it has merely been handed a string of letters and punctuation marks. As a result of this, no ideas will be created, no knowledge will be consulted, no imagery will be formed. Yet simply because highly evocative English words are embedded in the string, it is hard to resist this easy slide down a very slippery epistemological slope.

6. Among the complex real-world analogies they cite it as having made are a political one (involving alleged terrorists in Nicaragua, Hungary, and Israel), numerous scientific analogies (including the water-flow and heat-flow case), a number of "jealous animal stories", and some that link vastly different domains of knowledge. In fact, giving their program the ability to make cross-domain analogies is clearly one of the achievements of which Holyoak and Thagard are most proud.

7. Do even Holyoak and Thagard themselves understand exactly how empty their symbols are? Let as assume they do. But in that case, what I don't understand at all is how they could feel comfortable in claiming that ACME carries out "cross-domain analogy-making", when in fact ACME deals with no domains at all – just strings.

8. Clearly, all AI researchers, myself included, want to brag about their programs' achievements; on the other hand, we all know that we can't get away with out-and-out anthropomorphism. What generally results is some kind of intermediate level of description, in which a bit of caution is used but much is left ambiguous, so that readers are still free to draw conclusions that often will amount to some kind of Eliza effect - benefiting the researchers, needless to say. It may well be that in this book, precisely this kind of thing takes place in our discussions of our own work, but there is one difference: our domains are deliberately so stripped-down that the claims made cannot be very grandiose.