Joshua Harkness

CSC 366

Prof. Graci

11/9/17

The Ineradicable Eliza Effect and Its Dangers Review

1. There is an insidious problem in writing about such a computer achievement, however. When someone writes or reads "the program makes an 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.
2. Needless to say, it turns out that the program in question knows none of these facts. Indeed, it has no concepts, no permanent knowledge about anything at all. For each separate analogy it makes (it is hard to avoid using that phrase, even though it is too charitable), it is simply handed a short list of "assertions" such as "Liquid (water) ", "Greater (Pressure (beaker) , Pressure (vial) )", and so on.
3. But the computer doesn't care at all that this makes no sense, because it is 54not reaching back into a storehouse of knowledge to relate the words in these assertions to anything else. The terms are just empty tokens that have the form of English words.
4. There is an irresistible tendency to conflate the rich imagery evoked by the drawings with the computer data-structures printed just below them (FigureVI-2, page 277). For us humans, after all, the two representations feel very similar in content, and so one unwittingly falls into saying and writing 'The computer made an analogy between this situation and that situation." How else would one say it?
5. 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 - especially 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.
6. When one really looks closely at what has been done, the achievement typically shrinks and shrinks until one sees that there was very little knowledge of the real-world concepts purported to have been manipulated, and that moreover, exactly the proper concepts (in an unimaginably diluted form) were supplied, and few others.
7. Such close scrutiny, carried out regrettably seldom by AI researchers, forces one to confront in great depth questions about where and when meaning is present - questions about how and when meanings are truly carried by symbols.
8. Obviously- so the argument would run – Holyoak and Thagard were simply avoiding a whole series of complex and awkward turns of phrase by speaking as if the terms in the expressions handed to ACME really referred to real-world entities and relationships - and surely this is a useful and harmless move to make. Surely nobody would misunderstand it. Ah, but there's the

rub. People, even sophisticates like Boden and Waldrop, make misunderstandings of this sort all the time

9. 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.
10. Whereas many research groups appear to be tackling domains of such complexity that even human experts are cowed- thermodynamics, international terrorism, atomic physics, computer-system configuration, VLSI chip design, economic forecasting, and on and on – we in our group are dealing with such microscopic domains that our programs' achievements appear to be nearly trivial. When there's a little kid trying somersaults out for the first time next to a flashy gymnast doing flawless flips on a balance beam, who's going to pay any attention to the kid?. Our projects in microdomains come across, at least on the surface, as the equivalent of the little kid doing a somersault. Even the name "Copycat" was deliberately chosen to downplay the program's level of expertise - to emphasize its childlikeness.