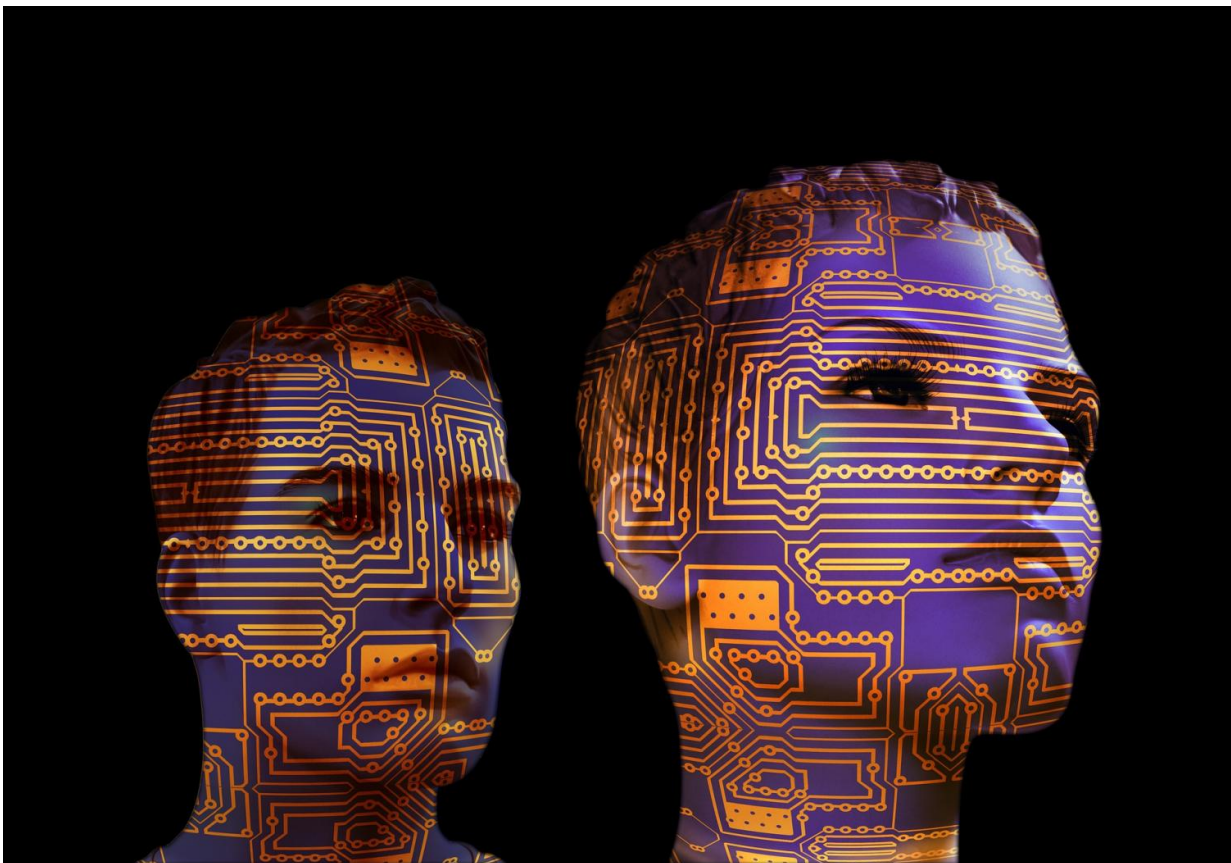# Chatbots need to smarten up but easier said than done

July 15 2016, by Nancy Owano

Credit: CC0 Public Domain

(Tech Xplore)—AI chatbots on the same level as humans? Nah. They still cannot impress when it comes to ambiguous sentences.

Graeme Burton, *The Inquirer*, on Friday reported on the Winograd Schema Challenge. This is a contest that seeks to reward "technologists who can build a system that understands the kind of ambiguous sentences humans come out with all the time, but which are simple for other humans, even stupid ones, to understand."

Even the very best two entrants at the event got correct scores only 48 per cent of the time, Burton said.

Will Knight in *MIT Technology Review* elaborated on the test results: "The best two entrants were correct 48 percent of the time, compared to 45 percent if the answers are chosen at random. To be eligible to claim the grand prize of $25,000, entrants would need to achieve at least 90 percent accuracy."

Some examples of what we would easily get but would mess up the chatbot? Here are two such examples of what computers would have to make sense out of; the sentences are ambiguous but simple for humans to work out as the latter can 'read' context. This is from the Commonsense Reasoning ~ Winograd Schema Challenge page:

I. The trophy would not fit in the brown suitcase because it was too big (small). What was too big (small)?
Answer 0: the trophy
Answer 1: the suitcase

II. The town councilors refused to give the demonstrators a permit because they feared (advocated) violence. Who feared (advocated) violence?
Answer 0: the town councilors
Answer 1: the demonstrators

Burton referred to what Gary Marcus, a research psychologist at New

York University, said was the machine problem (Marcus was an advisor for the event)—namely, said Burton, "computers lack common sense, and programming it into them is incredibly difficult."

Wait, isn't this like the Turing test, where we got to enjoy Eugene? No, this is different, said Burton. "The Challenge is deliberately designed to be different from the Turing Test, which tests only whether a human can be fooled into thinking that an AI program is human."

He said Marcus felt the language test was a more objective test of genuine AI. (Burton remarked that, with the Turing Test, "there are more than enough idiots who could be fooled into helping an AI system to pass that test.")

The Commonsense Reasoning ~ Winograd Schema Challenge page similarly made the observation that "Chatbots like Eugene Goostman can fool at least some judges into thinking it is human, but that likely reveals more about how easy it is to fool some humans, especially in the course of a short conversation, than the bot's [intelligence](.)."

Knight in *MIT Technology Review* said the results of the contest were presented recently at an academic conference in New York. Knight also provided a background history of the challenge.

"Winograd Schema sentences were first highlighted as a way to gauge machine comprehension by Hector Levesque, an artificial-intelligence researcher at the University of Toronto. They are named after Terry Winograd, a pioneer in the field and a professor at Stanford University who built one of the first conversational computer programs."

Knight said the challenge was proposed in 2014 as an improvement on the Turing Test.