Language as a Litmus Test

Language, which clearly played an important role in human evolution, has long been considered a hallmark of human intelligence, and when Barbara Grosz started working on problems in artificial intelligence (AI) in the 1970s, it was the litmus test for defining machine intelligence. The idea that language could be used as a kind of Occam's razor for identifying intelligent computers dates to 1950, when Alan Turing, the British scientist who cracked Nazi Germany's encrypted military communications, suggested that the ability to carry on a conversation in a manner indistinguishable from a human could be used as a proxy for intelligence. Turing raised the idea as a philosophical question, because intelligence is difficult to define, but his proposal was soon memorialized as the Turing test. Whether it is a reasonable test of intelligence is debatable. Regardless, Grosz says that even the most advanced, language-capable AI systems now available—Siri, Alexa, and Google—fail to pass it.

The Higgins professor of natural sciences has witnessed a transformation of her field. For decades, computers lacked the power, speed, and storage capacity to drive neural networks—modeled on the wiring of the human brain—that are able to *learn* from processing vast quantities of data. Grosz's early language work therefore involved developing formal models and algorithms to create a computational model of discourse: telling the computer, in effect, how to interpret and create speech and text. Her research has led to the development of frameworks for handling the unpredictable nature of human communication, for modeling one-on-one human-computer interactions, and for advancing the integration of AI systems into human teams.

The current ascendant AI approach—based on neural networks that learn—relies instead on computers' ability to sample vast quantities of data. In the case of language, for example, a neural network can sample a corpus—extending even to everything ever written that's been posted online—to learn the "meaning" of words and their relationship to each other. A dictionary created using this approach, explains assistant professor of computer science Alexander "Sasha" Rush, contains mathematical representations of words, rather than language-based definitions. Each word is a vector—a relativistic definition of a word in relation to other words. Thus the vectors describing the relationship between the words "man" and "woman" would be mathematically analogous to those describing the relationship between words such as "king" and "queen."

This approach to teaching language to computers has tremendous potential for translation services, for developing miniaturized chips that would allow voice control of all sorts of devices, and even for creating AIs that could write a story about a sporting event based purely on data. But because it captures all the human biases associated with culturally freighted words like "man" and "woman," and what the ensuing mathematical representations might embody with respect to gender, power dynamics, and inequality when confronted with the associations of a word such as "CEO," it can lead neural-network based AI systems to produce biased results.

Rush considers his work—developing language capabilities for microscopic computer chips—to be purely engineering, and his translation work to be functional, not literary, even though the goal of developing an AI that can pass the Turing test is undoubtedly being advanced by work like his. But significant obstacles remain.

How can a computer be taught to recognize inflection, or the rising tone of words that form a question, or an interruption to discipline kids ("Hey, stop that!"), of the sort that humans understand immediately? These are the kinds of theoretical problems Grosz has been grappling with for years. And although she is agnostic about whatever approach will ultimately succeed in building systems able to participate in everyday human dialogue, probably decades hence, she does allow that it might well have to be a hybrid of neural-network learning and human-developed models and rules for understanding language in all its complexity.

more quickly. When such fundamentals of intelligent-systems design aren't respected, the systems are assumed to be capable of things they can't do, or are used in naïve, inappropriate ways.

Grosz's highly interdisciplinary approach to *research*, informed by linguistics, philosophy, psychology, economics, and even a bit of anthropology and sociology, led her to think also about which of these subjects might best inform the *teaching* of AI systems design. Though she had taught an introductory course on AI from 1987 to 2001, a time when its application remained largely theoretical, the world had changed by the time she rebooted that course in 2013 and 2014, when fully operational AI systems were being deployed. Grosz realized there was a teaching opportunity in the interplay between the ethical challenges presented by AI and good systems design.

This led to one of Grosz's most important contributions to the teaching of computer science at Harvard: the idea that ethics should be tightly integrated into every course. In the fall of 2015, she introduced a new course, "Intelligent Systems Design and Ethical Challenges." By the following year, more than 140 students had applied for the 25 spots in the class, emboldening her to encourage her computer-science colleagues to incorporate some teaching of ethics into their own courses. Because most of them lacked sufficient background to be comfortable teaching ethics, she began a collaboration with Wolcott professor of philosophy Alison Simmons, who chairs the philosophy department. Together, they worked with colleagues in their respective fields, enlisting CS professors willing to include ethics modules in their computer-science courses and philosophy graduate students to teach them.

The aim of this "Embedded EthiCS" initiative, she says, is to instruct the people who will build future AI systems in how to identify and think through ethical questions. (Computer science is now the second largest concentration among Harvard undergraduates; if students from related fields such as statistics or applied mathematics are included, the total enrollment substantially exceeds that of top-ranked economics.) "Most of these ethical challenges have no single right answer," she points out, "so just as [the students] learn fundamental computing skills, I wanted them to learn fundamental ethical-reasoning skills." In the spring of 2017, four computer-science courses included some study of ethics. That fall, there were five,