Behavioral estimates of conceptual structure are robust across tasks in humans but not large language models

Siddharth Suresh
University of Wisconsin-Madison, Madison, Wisconsin, United States

Kushin Mukherjee
University of Wisconsin-Madison, Madison, Wisconsin, United States

Lisa Padua
Albany State University, Albany, Georgia, United States

Timothy Rogers
University of Wisconsin- Madison, Madison, Wisconsin, United States

Abstract

Neural network models of language have long been used as a tool for developing hypotheses about conceptual representation in the mind and brain. For many years, such use involved extracting vector-space representations of words and using distances among these to predict or understand human behavior in various semantic tasks. In contemporary language AIs, however, it is possible to interrogate the latent structure of conceptual representations using methods nearly identical to those commonly used with human participants. The current work uses two common techniques borrowed from cognitive psychology to estimate and compare lexical-semantic structure in both humans and a well-known AI, the DaVinci variant of GPT-3. In humans, we show that conceptual structure is robust to differences in culture, language, and method of estimation. Structures estimated from AI behavior, while individually fairly consistent with those estimated from human behavior, depend much more upon the particular task used to generate behavior responses–responses generated by the very same model in the two tasks yield estimates of conceptual structure that cohere less with one another than do human structure estimates. The results suggest one important way that knowledge inhering in contemporary AIs can differ from human cognition.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.