Artificial Neural Networks Reveal a Cognitive Continuum Toward Human Abstraction

Li Wenjie
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Margaret Henderson
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Yonatan Bisk
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Jessica Cantlon
Carnegie Mellon, Pittsburgh, Pennsylvania, United States

Abstract

Do neural network models that fail to behave human-like reflect a fundamental divergence from human cognition, or do they mirror earlier developmental or evolutionary stages? We propose that such models may, in fact, offer insights into the origins of human abstraction. We evaluated over 200 pretrained neural networks alongside macaques, Tsimane natives, US adults and children on three visual match-to-sample tasks targeting increasing levels of abstraction: visual-semantic similarity, shape regularity, and relational reasoning. As task demands grow more abstract, just like monkey's, model decisions increasingly diverge from adult human behavior. However, representational similarity analyses reveal shared internal structure with all human groups, suggesting overlapping cognitive strategy. We further show that model alignment depends on specific design choices—architecture, scale, training regime, and language supervision—highlighting which inductive biases support human-like abstraction.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*