The Uncanny Valley meets the Humorous Hill: Things are funny when they match a pattern but fall short on quality

Antara Bhattacharya
Harvard University, Cambridge, Massachusetts, United States

Jennifer Hu
Harvard University, Cambridge, Massachusetts, United States

Tomer D. Ullman
Harvard University, Cambridge, Massachusetts, United States

Abstract

We propose the "Humorous Hill" hypothesis: things are funny when they match an expected pattern, but fall short of being good. We suggest this form of humor underlies the amusement felt towards many of children's utterances, and much of the recent engagement with AI. We tested the Humorous Hill by using language models (LMs) to create novel examples of open-ended categories in two domains (paint colors and movie titles). The LMs varied in quality and architecture, including n-gram models with increasing windows (2-grams to 9-grams), and increasingly sophisticated transformer-based models (GPT babbage to GPT-4). Participants (N=300) rated items for category membership, goodness of fit, and humor. Across models and categories, we found an inverted U-shaped relationship between humor and accuracy. We propose that much of people's engagement with artificial agents is driven by finding their outputs humorous, rather than good – a form of humor that also applies to children.

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