You are what you're for: Essentialist categorization in large language models

Siying Zhang, Jingyuan S. She, Tobias Gerstenberg & David Rose

Department of Psychology, Stanford University

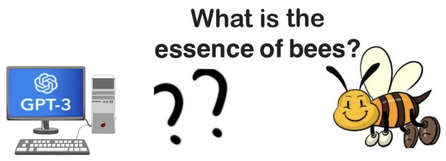
Introduction

Background:

Prior research has shown that humans engage in essentialist categorization, meaning that they categorize things based on their underlying properties, rather than appearance.

Question:

Do LLMs tend to categorize on the basis of essential properties or on the basis of described appearance?



Hypothesis:

LLMs are more likely to categorize things based on essential properties than on described appearance.

Approach:

- Show LLMs (OpenAI's GPT-3 and BigScience's BLOOM) vignettes from the literature on essentialist categorization.
- Examine whether in a classic test of essentialist categorization – the transformation task – LLMs prioritize essential properties over information about what something looks like.

Analysis of prior work

Methods:

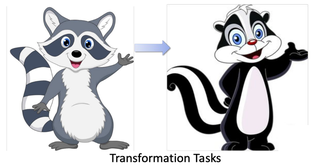
Investigated whether the outputs from LLMs match those of people on a set of experiments on essentialist thinking about categories → Replicated the studies from selected papers and queried GPT-3 & BLOOM.

Results:

- GPT-3's judgments were inconsistent with those of human participants in some of the studies. The exceptions were the studies that provided *teleological information* or information about *what the things were made of*.
- LLMs displayed a tendency to trace essential properties to determine category membership.

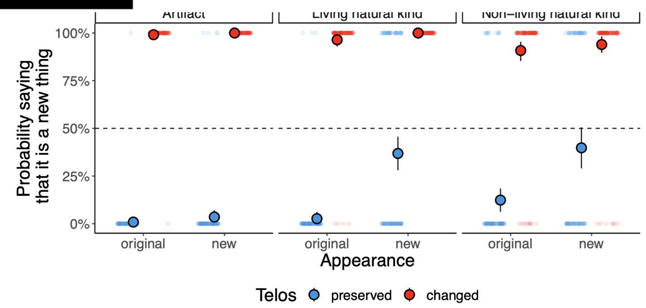
Experiment 1

Do teleological considerations play a role in LLM's categorization?



Some very talented and skilled scientists decide that they are going to perform a special procedure to turn lotion into a bed. After the special procedure, the thing looked like a bed. After running some tests, they found that the thing after the special procedure didn't provide a place to sleep. Instead, it only moisturized and softened the skin.

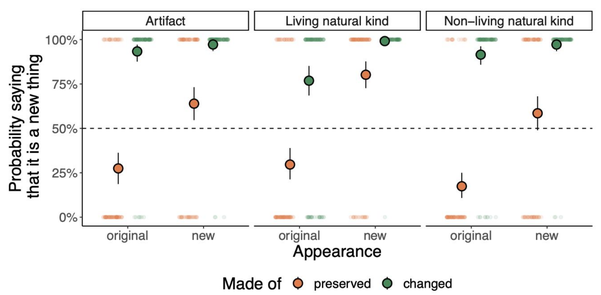
Is the thing after the procedure lotion or a bed?



Results: Teleological considerations carry more weight than appearance when categorizing things that change.

Experiment 2

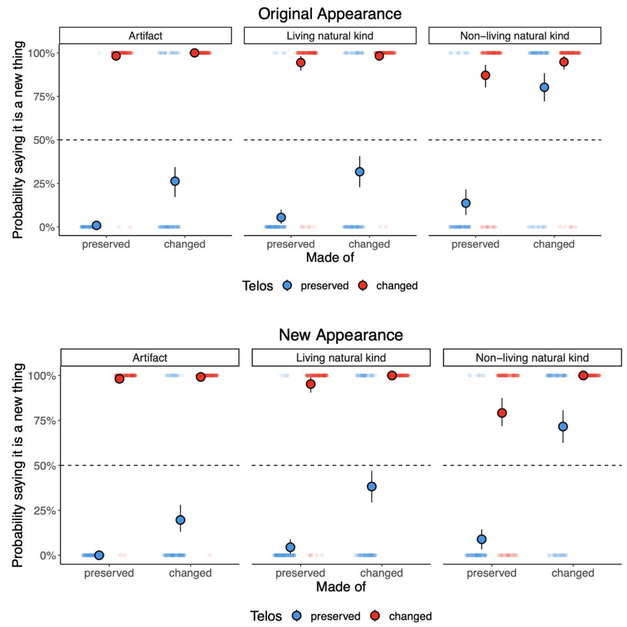
Does what something is made of play a role in LLM's categorization?



Results: What something is made of carries more weight than appearance when categorizing things that change.

Experiment 3

Does what something is made of or its telos matter more in LLM's categorization?



Results: Teleological considerations carry more weight than what something is made of or how it appears.

Discussion

GPT3 reflects a human bias toward teleological thinking

- ❏ Language suffices for transmitting essential beliefs. LLMs categorize based on essential properties.
- ❏ When comparing candidates for essential properties, what something is *for* matters more than what something is *made of.*
- ❏ Next step: What aspects of language are driving this?

Key References

- Gelman, S. A. (2003). The essential child: Origins of essentialism in everyday thought. Oxford Series in Cognitive Development.
- Keil, F. C. (1992). Concepts, kinds, and cognitive development. mit Press.
- Rose, D., & Nichols, S. (2019). Teleological essentialism. Cognitive Science, 43(4) & Rose, D., & Nichols, S. (2020). Teleological essentialism: generalized. Cognitive science, 44(3)