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

Siying Zhang* (syzhang6@stanford.edu), Department of Psychology, Stanford University

Jingyuan S. She* (jshe@haverford.edu), Haverford College

Tobias Gerstenberg (gerstenberg@stanford.edu), Department of Psychology, Stanford University

David Rose (davdrose@stanford.edu), Department of Psychology, Stanford University

*joint first authors

Abstract

How do essentialist beliefs about categories arise? We hypothesize that such beliefs are transmitted via language. We subject large language models (LLMs) to vignettes from the literature on essentialist categorization and find that they align well with people when the studies manipulated teleological information – information about what something is for. We examine whether in a classic test of essentialist categorization – the transformation task – LLMs prioritize teleological properties over information about what something looks like, or is made of. Experiments 1 and 2 find that telos and what something is made of matter more than appearance. Experiment 3 manipulates all three factors and finds that what something is *for* matters more than what it's *made of*. Overall, these studies suggest that language alone may be sufficient to give rise to essentialist beliefs, and that information about what something is for matters more.

Keywords: essentialism; large language models; teleology; categorization; transformation.

Introduction

If you're a deep neural network, then rotating a stop sign might make you categorize it as dumbbell or a ratchet (Heaven, 2019). A spider on a textured background might look to you like a manhole cover. And a mushroom jutting from the ground might appear to be a pretzel. But even if you are a human, you rely on visual information in categorization. Tell a kindergartner that a raccoon has undergone surgery so that it now looks like a skunk and they can't help but view the thing as a skunk (Keil, 1992). Ask if it would still be a skunk if it's mommy and daddy were raccoons, or if its babies were raccoons and you'll likely be met with a resounding "yes". Why would one say that a raccoon merely made to look like a skunk is now a skunk? In the words of one kindergartner: "Because it looks like a skunk, it smells like a skunk, it acts like a skunk, and it sounds like a skunk." (Keil, 1992, p. 188) While younger children tend to categorize things based on what they look like, older children and adults often go beyond mere appearance, and take into account more essential properties that really make a thing the thing that it is (Medin & Ortony, 1989; Gelman, 2003; Keil, 1992; Atran, 1995). Essential properties are those that give rise to a thing's identity and make something a member of a kind. They are the properties we trace and expect to persist over time, even if a thing's appearance changes radically (Kalish, 1995; Gelman, 2003; Rose & Nichols, 2019, 2020).

A great deal of evidence suggests that people categorize things on the basis of their essential properties (e.g., Gelman & Wellman, 1991; Gelman, 2003; Keil, 1992; Newman & Keil, 2008; Rose & Nichols, 2019, 2020). An important question is how we come to have essentialist beliefs about categories. One possibility is that language alone gives rise to essentialist thinking. How could we test this? Recent advances in computational linguistics and artificial intelligence may help. In particular, large language models (LLMs), due to their impressive ability of tracking linguistic co-occurrence patterns in vast amounts of data, may serve as a test bed for exploring whether essentialist beliefs might arise from language input.

In recent years, LLMs have achieved impressive performance on tasks like question-answering, sentence completion, and coherent article generation (Brown et al., 2020). Piantadosi & Hill (2022) argue that LLMs may capture fundamental aspects of meaning and Weir et al. (2020) find that LLMs effectively infer generic concepts given their associated properties. For instance, they correctly infer "bear" given "has fur", "is big", and "have claws". A number of studies have even found that LLMs perform much like humans in ethical (Jiang et al., 2021), abstract (Dasgupta et al., 2022), and logical reasoning (Kojima et al., 2022). That said, many have argued that there is a fundamental difference between benchmark successes and real reasoning abilities. For instance, LLMs fail in most symbolic reasoning tasks and succeed only on a context-dependent basis (Talmor et al., 2019), mimic but fall short of humans' inductive reasoning (Han et al., 2022), are unable to generalize well to reasoning tasks out of the training set (Zhang et al., 2022), and struggle to distinguish between impossible and unlikely real-world events (Kauf et al., 2022). Tasks involving reasoning seem to be especially challenging for LLMs. However, essentialist categorization need not involve reasoning. Language alone could fuel it and may be sufficient to form essentialist beliefs.

Our question is thus: Are LLMs more inclined to categorize on the basis of essential properties versus more superficial properties such as a thing's visual appearance?

Essential properties

What kinds of essential properties might be candidates for LLMs to draw on in categorization? Here we focus on two: 1) teleological properties – or what something *is for* – and 2) what something *is made of*.

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

Teleological properties Teleological thinking plays a central role in explanation (e.g., Bloom, 2007; Kelemen, 1999; Kelemen & Rosset, 2009; Kelemen et al., 2013; Lombrozo & Carey, 2006; Lombrozo et al., 2007; Lombrozo & Rehder, 2012; Foster-Hanson & Lombrozo, 2022). We often answer 'why' questions by pointing out goals and purposes. Teleological thinking also plays a role in people's causal judgments (Rose & Schaffer, 2017) and even matters for people's judgments about whether an object exists, that is, whether some collection of parts composes a whole object, and persists through changes over time (Rose, 2015; Rose & Schaffer, 2017; Rose, 2019, 2020; Rose et al., 2020). Rose (2020) suggests that the central role of teleology in explanation and judgements of existence and persistence might arise because we essentialize categories in terms of teleology.

Utilizing classic tests of essentialist thinking, Rose & Nichols (2019, 2020) provide support for the claim that people essentialize a broad range of categories – from artifacts to non-living natural kinds – in terms of teleology. To take just one example, people think that bees are for making honey and spiders are for spinning webs (Rose & Nichols, 2019). And if a bee undergoes radical transformation so that it now looks like a spider, but preserves the bee telos – that is, it still makes honey – people categorize the creature as a bee. But if it instead changes its telos after the transformation to that of a spider – spinning webs – people say it's a spider. It seems that we trace the persistence of a thing's telos across change. This suggests that teleological properties are treated as essential properties.

What something is made of One important view of essentialism, one that contrasts with teleological essentialism and has dominated the psychological literature for well over three decades, maintains that what something is made of, what it is constituted by, determines its essence (Gelman, 2003; Keil, 1992). This view is inspired by Kripke (1972) and Putnam (1962), who claim that, for instance, the essence of water is H_2O. This example, that H_20 is the essence of water, is a leading example that animates the view that so called "scientific" properties are the kind of properties that serve as essential properties. It might be that what something is made of plays an important role in LLM's essentialist categorization judgments.

Essentialized categories These two views not only disagree on what properties are essentialized but also on what categories are essentialized. Those who maintain that essential properties are associated with what something is made of also maintain that only living natural kinds, such as racoons or kangaroos, and non-living natural kinds, such as rocks and lightning, are essentialized (e.g., Gelman, 2003; Keil, 1992). Artifacts, on this view, are not essentialized. Those who maintain that teleological properties are treated as essential properties, hold that living and non-living natural kinds as well as artifacts, such as clocks and hotplates, are essentialized (e.g., Rose & Nichols, 2020).

Do LLMs engage in essentialist categorization?

Our question is whether LLMs are more inclined to categorize on the basis of essential properties than on the basis of described appearance. To test this we focus on two main tasks: 'transformation tasks' and 'nature vs. nurture tasks'. Transformation tasks involve determining whether something persists after undergoing changes in properties (e.g., Keil, 1992). Nature vs. nurture tasks involve determining whether an animal raised by a different animal would be a member of the new animal category (e.g., Gelman & Wellman, 1991). We will explore LLMs' categorization judgments on a range of domains, from living and non-living natural kinds to artifacts.

The language models we use are OpenAI's GPT-3 (Brown et al., 2020) and BigScience's BLOOM (Scao et al., 2022). GPT-3 is one of the best-performing general-purpose language models; BLOOM is the largest open multilingual language model. We begin by testing whether GPT-3 and BLOOM categorize things like people do in a set of studies that have investigated essentialist thinking.

Analysis of Prior Work

Please find links to the pre-registrations, data, and analyses files here: `https://github.com/cicl-stanford/essentialism_in_llms`. The goal of this study was to investigate whether the outputs from LLMs match those of people on a set of experiments aimed at documenting essentialist thinking about categories.

Methods

Materials We ran studies from Rose & Nichols (2019, 2020); Gelman & Wellman (1991); Keil (1992); Waxman et al. (2007); Hampton et al. (2007); Barton & Komatsu (1989) on GPT-3 and BLOOM. These studies were selected because they are representative of work that has used 'transformation tasks' and 'nature vs. nurture tasks' to investigate essentialist categorization across different domains. We also chose these studies because they're mostly text and relied little, if at all, on visual stimuli.

Design and Procedure We replicated the studies from the selected papers as closely as possible. We followed the same procedure and presented all materials from each experiment, one by one, to both GPT-3 (Model: `text-curie-001`) and BLOOM. For each experimental condition, we queried the two language models 50 times.

Data Processing GPT-3 and BLOOM return open ended responses. To process these responses, we trained GPT-3 to retrieve single-word responses from the full text responses. To do so, we gave GPT-3 a small selection of sentences with correct item names as training data as part of the prompt. For each experiment condition (129 in total), we manually selected 3 sentences that we judged to be easy (e.g., "A tire."), normal (e.g.,"In this scenario, the doctors would have a tire."), and difficult (e.g., "If you consider the act of cutting and sewing the tire to be the operation, then the doctors ended up

Table 1: GPT-3 and BLOOM's degree of alignment with participants' categorization judgments in prior work. T = transformation tasks, N = nature vs. nurture tasks. We marked each model-generated answer as correct or incorrect based on the majority choice in the paper. We calculated accuracy for each study by averaging over all 50 responses generated by the LLMs.

Paper	RN2019	RN2019	RN2020	RN2020	GW1991	GW1991	Keil1992	Waxman2007	Hampton2007	BK1989
Task	T	N	T	N	T	N	T	N	T	T
# of studies	2	2	2	1	1	2	1	3	1	1
GPT-3 accuracy	75%	69%	63%	70%	37%	19%	49%	7%	91%	72%
BLOOM accuracy	48%	48%	59%	65%	75%	40%	42%	27%	68%	2%

with a boot. Doctors would see the operation as a success."). We paired these sentences with the desired single word that should be retrieved (e.g., "tire"). These training data were appended to the beginning of the query so that GPT-3 would refer to it and deliver retrieved answers to the new sentences that were fed in. To test that GPT-3 was accurate at extracting responses, we randomly sampled half of the responses and have two of the authors checked. GPT-3 performed this single-word extraction task with 99% accuracy. Responses that weren't clear, such as, "The things that hatch from the eggs will be neither bees nor spiders", were coded as "unsure".

Results and Discussion

Table 1 shows that GPT-3's judgments were inconsistent with those of human participants in some of the studies. The exceptions were the studies by Rose & Nichols (2019, 2020), Hampton et al. (2007) and Barton & Komatsu (1989). Here GPT-3's judgments were closely aligned with those of humans. The main difference between the studies where GPT-3 gave judgments that were inconsistent with humans and those consistent with humans, is that in the cases where GPT-3 gave consistent judgments teleological information (Rose & Nichols, 2019, 2020; Barton & Komatsu, 1989) or information about what the things were made of (Barton & Komatsu, 1989) was included. The work by Hampton et al. (2007) showed very high agreement. This work involved cases where an animal changed and now looks and acts like a different animal. Both LLMs and humans categorize the thing as a different animal in that case. But this shouldn't be surprising. Indeed, this should be viewed as a control given that the animals underwent total transformation.

Experiment 1: Telos vs. Appearance

We found that LLMs made judgments much like people in cases where teleological information was provided. In this study, we continue to investigate what role teleology plays in LLM categorization in transformation tasks. We generated a fresh set of pairs of items from each domain and adapted and simplified the transformation vignettes from (Rose & Nichols, 2019, 2020). Below is the transformation task vignette from Rose & Nichols's (2019) study 1:

> Some very talented and skilled scientists, Suzy and Andy, decide that they are going to perform a special operation on a bee. They removed its wings and antennae, lengthened its legs and added a new pair of legs. They also inserted into the back of it something for making webs and trained the animal so that it would eat insects.

On the same page, participants were then shown images of a bee and a spider, indicating the thing before and after the operation. Participants were then told:

> After running some test, they found that the thing after the special operation didn't [original thing's telos / new thing's telos]. Instead, it only [new thing's telos / original thing's telos].

Then they were asked: "To what extent do you think that the thing after the special operation is [an original thing / a new thing]?"

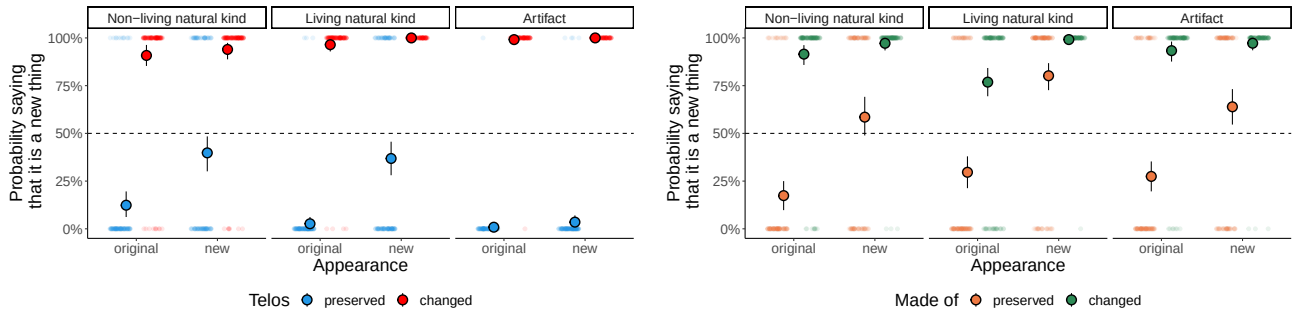
Methods

Materials We generated a list of items that we expected to have mutually exclusive purposes. Four items were selected for each domain. *Non-living natural kinds*: lightning, cloud, sun, and soil; *Living natural kinds*: chicken, cow, worm, and bat; *Artifacts*: bed, microwave, lotion, and keychain. Then we queried GPT-3 about each item's telos ten times by asking, for example, "What is the purpose of a chicken?". We selected the top mentioned answer (e.g., "A chicken's purpose is to produce eggs.") as each item's telos in the subsequent experiment. We used the following template as a prompt:

> Some very talented and skilled scientists decide that they are going to perform a special procedure to turn a/an [original thing] into a/an [new thing]. After the special procedure, the thing looked like a/an [original thing / new thing]. After running some tests, they found that the thing after the special procedure didn't [original thing's telos / new thing's telos]. Instead, it only [new thing's telos / original thing's telos].

Design Our design was a 2 (Appearance: original, new) $\times$ 2 (Telos: preserved, changed) $\times$ 3 (Domain: living natural kind, non-living natural kind, artifact) design.

To control for potential order effects, we counterbalanced the order of telos and appearance information in the prompt. We also counterbalanced which item was turned into what. There were 12 possible item pairs in each domain. GPT-3 and BLOOM received all item pairs within a domain. We elicited 5 responses for each pair by condition.



(a) GPT-3 Telos vs. Appearance

(b) GPT-3 Insides vs. Appearance

Figure 1: **Experiment 1 and 2.** Categorization ratings across all three domains in telos preserved (blue) and telos changed (red) conditions(left) and insides preserved (orange) and insides changed (green) conditions(right). The x-axis indicates whether the appearance of the thing after the special operation was the same (original) or changed (new).

We used the following model setting in this and all following experiments. For GPT-3, we used the text-davinci-002 model, set the maximum output tokens to be 50, and used the default setting on everything else (Temperature = 0.7, maximum length = 256, Top p = 1, Frequency penalty = 0, Presence penalty = 0, Best of = 1, Show probabilities = off). Although GPT-3 is able to show probabilities, we decided not to use this feature because GPT-3 shows probabilities by tokens not by words. This makes it complicated to compute the probability for single words. For BLOOM, we set the maximum output token to 5, the temperature to 0.7, the length penalty to 1.0 and set the random seed to 2022.

Procedure GPT-3 and BLOOM were given a case that involved a thing undergoing a special procedure so that it either had the same or different appearance and had either the original thing's telos or a different telos. After the vignette, the LLMs were asked: "Is the thing after the special operation a [original thing] or a [new thing]?"

Here is an example:

> Some very talented and skilled scientists decide that they are going to perform a special procedure to turn a chicken into a worm. After the special procedure, the thing looked like a chicken. After running some tests, they found that the thing after the special procedure didn't produce eggs. Instead, it only helped decompose organic matter. Is the thing after the special operation a chicken or a worm?

Data Processing For the current and all the following studies, we had two independent coders manually extract item names from the full responses. To guarantee accuracy, we asked the coders to check each other's work and discuss any discrepancies. For all the unresolved discrepancies, we assigned "unsure" as the response.

Results

Figure 1a shows the proportion of cases in which the model said that the transformed thing was a new thing as a function of appearance and telos, separately for each domain. Throughout we analyzed the models' responses by running Bayesian logistic regressions using the `brms` (Bürkner, 2017) package in `R` (R Core Team, 2019). We used the `emmeans` package (Lenth et al., 2019) to analyze differences between conditions. We report the medians and 95% credible intervals of the posterior distributions and call something an effect when the credible interval excludes 0. Here in Experiment 1, we fit a Bayesian logistic regression model with telos, appearance, and their interaction as predictors.

As Table 2 shows, GPT-3 was more inclined to judge that something was a new thing after the transformation when its telos had changed compared to when its telos was preserved. Likewise, for appearance. When something changed its appearance, GPT-3 was more inclined to judge that it changed categories. The effect of teleology was greater than the effect of appearance. And it was greater for artifacts than living or non-living natural kinds, and for living natural kinds than non-living natural kinds.

BLOOM displayed a much weaker and more mixed pattern. It was also more inclined to judge that something was a new thing when its telos changed. But in contrast to GPT-3, changes in appearance didn't affect BLOOM's categorization

Table 2: **Experiment 1**: Posterior distributions of the overall difference of telos, the difference of telos across domains, the difference in appearance and the difference between telos and appearance for GPT-3 and BLOOM. The values show medians with 95% credible intervals in brackets.

	GPT-3	**Bloom**
Telos	.80[.77, .83]	.15[.11, .21]
artifacts - living natural kinds	.19[.12, .25]	.04[−.06, .16]
artifacts - non-living natural kinds	.31[.24, .37]	.04[−.07, .16]
living natural kinds - non-living natural kinds	.12[.05, .18]	.0[−.11, .11]
Appearance	.11[.09, .14]	.06[.00, .10]
Telos - Appearance	.68[.64, .72]	.09[.03, .16]

judgments. The effect of teleology was greater than the effect of appearance. But the effect of teleology was not greater for any domain.

Discussion

Teleological considerations play a role in LLMs categorization judgments when considering things that undergo radical transformation. This was clearly the case for GPT-3 and less so for Bloom. Transformation tasks provide some of the best evidence of essentialist thinking. And given that GPT-3 heavily relies on teleological considerations when judging category membership across transformation, this suggests that language models, like GPT-3, might be more inclined to take into account essential properties, in this case, teleological properties, than appearance in categorization. And interestingly, though teleological considerations play the most powerful role in GPT-3's judgments about artifacts undergoing transformation, they also affect the categorization of non-living and living natural kinds. Across a range of things, GPT-3 places heavy weight on teleological considerations.

Experiment 2: Insides vs. Appearance

Experiment 1 indicates that teleological considerations play a role in essentialist categorization. But another view of essential properties has it that what something is made of, what it is constituted by, determines its essence. In this pre-registered study, we examine how LLMs respond when what something is made of changes or is preserved after transformation.

Methods

Materials We used the same list of items as in Experiment 1. We queried GPT-3 about what each item was made of (e.g., according to GPT-3, "Lightning is made of electrons, protons and other charged particles.") and then used that to vary what the things were made of in our experiment. We used the same vignette template but replaced telos information with "made of" information.

Design and Procedure The design and the procedure were the same as in Experiment 1.

Results

Figure 1b shows the LLMs' categorizations as a function of appearance and whether what it was made of was preserved or changed, separately for the three different domains. The results in Table 3 show that GPT-3 was more inclined to judge that something that changed what it was made of was now a new thing. Appearance, again, also mattered. However, what something was made of had a greater effect on categorization than appearance. The effect of what something was made of was greater for artifacts than living natural kinds but no different between artifacts and non-living natural kinds. The effect was smaller for living kinds than non-living natural kinds.

BLOOM, as in Experiment 1, displayed a much weaker and more mixed pattern. It judged that something that changed what it was made of changed categories. Appearance had no effect. What something was made of didn't have a greater effect on categorization than appearance and what something was made of was no different across domains.

Table 3: **Experiment 2**: Posterior distributions of the overall difference of "made of", the difference of "made of" across domains, the difference in appearance and the difference between "made of" and appearance for GPT-3 and BLOOM. The values show medians with 95% credible intervals in brackets.

	GPT3	**Bloom**
Made of	$.46[.42, .50]$	$.08[.03, .12]$
artifacts - living natural kinds	$.17[.08, .25]$	$.05[-.05, .16]$
artifacts - non-living natural kinds	$-.06[-.15, .02]$	$-.04[-.16, .05]$
living natural kinds - non-living natural kinds	$-.23[-.32, -.14]$	$-.11[-.22, .01]$
Appearance	$.26[.22, .30]$	$.04[.00, .09]$
Made of - Appearance	$.20[.14, .25]$	$.04[-.02, .10]$

Discussion

Experiment 1 showed that teleological considerations play an important role in LLMs' categorization. Here, we found that what something is made of also matters. In our final experiment, we directly pit appearance, telos, and what something is made of against one another to see what factor most strongly affects LLMs' categorizations.

Experiment 3: Telos, Insides & Appearance

In this last study, we combined both telos and "made of" information in the prompts to test whether LLMs favor one type of information over the other when determining category membership. Our experiment varies whether a thing has preserved or changed appearance, preserved or changed what it is made of, and preserved or changed its telos across transformations.

Methods

Materials We continued to use all items that were included in Experiment 1 and 2. The prompts were extended by including both telos and 'made of' information.

Design and Procedure Our design was a 2 (Appearance: original, new) $\times$ 2 (Telos: preserved, changed) $\times$ 2 (Made of: preserved, changed) $\times$ 3 (Domain: living natural kind, non-living natural kind, artifact) design. The procedure was the same as in Experiment 1 and 2.

Results and Discussion

Figure 2 shows GTP-3's categorizations as a function of the manipulated factors. As Table 4) shows, we found for GPT-3 that across all domains, the effect of teleology was greater than the effect of what something was made of. And the effect of what something was made of was greater than appearance. BLOOM again presented a weaker, more mixed set of results. The effect of teleology was only greater than what something was made of for artifacts and no different for the other kinds. The effect of what something was made of was greater than appearance for both artifacts and non-living natural kinds.

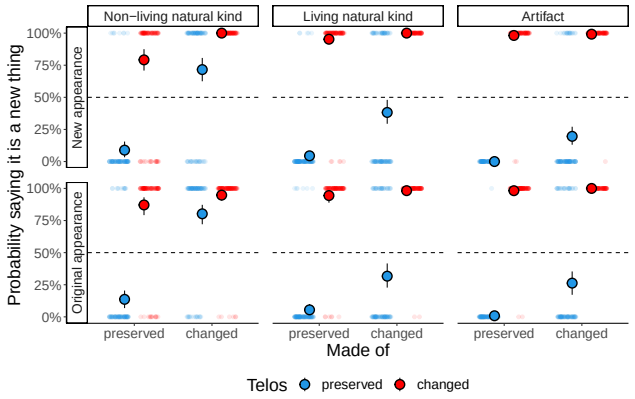


Figure 2: **Experiment 3**: GPT-3 categorization ratings across telos (preserved/changed) and made of (preserved/changed) conditions, separated by domain (columns) and appearance type (rows).

General Discussion

People don't categorize things based only on their appearance. They also take into account essential properties. We suggested that language alone may be sufficient to communicate such essential properties. We tested this by asking whether LLMs, such as GPT-3 and BLOOM, categorize on the basis of essential properties versus on the basis of a thing's described appearance.

We focused on two candidate essential properties: teleological properties and what something is made of. In Experiment 1, we found that teleological considerations played a greater role than appearance in GPT-3's categorization judgments. We also found, in Experiment 2, that what something is made of played a greater role than what it looked like. Experiment 3 pitted all factors against one another. We found that teleological considerations carried more weight than what something was made of, and that what something was made of mattered more than its appearance. This suggests that GPT-3 prioritizes teleological properties in essentialist categorization.

BLOOM delivered judgments that were weaker and more varied than GPT-3. This isn't surprising. GPT-3 is state-of-art and BLOOM, while having the virtue of being open source, is much more unwieldy. For instance, in querying BLOOM, one needs to set serious restrictions on it to prevent it from generating rambling, borderline incoherent, book length responses. In doing so, most of its responses tend to be "unsure". Thus, the results from BLOOM should be interpreted with caution.

Table 4: **Experiment 3**: Posterior distributions of the difference of telos effect and "made of" and the difference in "made of" and appearance across domains. The values show medians with 95% credible intervals in brackets.

	GPT-3	**Bloom**
Telos - Made of		
artifacts	$.74[.70, .80]$	$.12[.04, .20]$
living natural kinds	$.59[.54, .64]$	$.06[-.01, .14]$
non-living natural kinds	$.07[.02, .13]$	$.05[-.03, .12]$
Made of - appearance		
artifacts	$.14[.09, .19]$	$.17[.09, .25]$
living natural kinds	$.14[.09, .19]$	$.13[.05, .20]$
non-living natural kinds	$.43[.37, .48]$	$-.05[-.13, .03]$

In GPT-3, teleological information strongly affects how it categorizes things. This provides evidence supporting the idea that language itself may be sufficient for transmitting essentialist beliefs, and that information about what something is for may be particularly important. The fact that what something is for mattered more than what it's made of when both are directly pitted against one another also bears on a recent objection raised about teleological essentialism (Neufeld, 2021). Teleology, as the objection goes, only serves as a cue about insides, what something is made of, and it is this, not teleology, that ultimately matters in categorization (see also Joo & Yousif, 2022). However, we found that, at least for LLMs, teleology carries more weight than what something is made of even when both factors are directly pitted against one another. Moreover, in our vignettes, we explicitly stated what a thing's telos was and what it was made of. This way, the concern that changes in a thing's telos might affect inferences about what its made of doesn't apply in our case. We found overall that appearance had a weaker influence on categorization than information about telos and what the thing was made of. This effect is particularly striking since, for appearance, we directly stated what it looked like by using the category label (e.g. "After the special procedure, the thing looked like a chicken.") rather than only describing the differences in visual appearance (e.g. "the thing has feathers and a beak.").

The claim that LLMs categorize based on essential properties is surprising. It might seem instead that mere diagnosticity plays a role in LLMs categorization. But there is good reason for maintaining that LLMs are categorizing based on essences. Transformations are the classic, and arguably best, test of essentialist thinking. Judging that something persists across radical change provides good evidence that we treat the persistence of those properties as essential. LLMs do this. So this suggests they categorize on the basis of essences. Still the fact that LLMs categorize based on essential properties suggests that language alone transmits essentialist beliefs. Of course, this doesn't tell us which aspects of language lead to essentialism. Our future work aims to address this.

Conclusion

Language serves as an important vehicle for how beliefs about the world are transmitted. When LLMs are asked to categorize things, they don't just care about what it looks like, they care about essential properties, too. And when different candidates for essential properties are pitted against one another, we find that what something is for matters more than what it's made of. The language we learn appears to treat, and favor, teleological properties as essential properties.

Acknowledgments

Tobias Gerstenberg was supported by a research grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI). David Rose was supported by a Stanford Interdisciplinary Graduate Fellowship.

References

Atran, S. (1995). Causal constraints on categories and categorical constraints on biological reasoning across cultures. *Causal cognition: A multidisciplinary debate (pp. 205–233)*..

Barton, M. E., & Komatsu, L. K. (1989). Defining features of natural kinds and artifacts. *Journal of Psycholinguistic Research*, *18*(5), 433–447.

Bloom, P. (2007). Religion is natural. *Developmental science*, *10*(1), 147–151.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). Language models show human-like content effects on reasoning. *ArXiv*, *abs/2207.07051*.

Foster-Hanson, E., & Lombrozo, T. (2022). What are men and mothers for? the causes and consequences of functional reasoning about social categories. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford Series in Cognitive Development.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition*, *38*(3), 213–244.

Hampton, J. A., Estes, Z., & Simmons, S. (2007). Metamorphosis: Essence, appearance, and behavior in the categorization of natural kinds. *Memory & Cognition*, *35*(7), 1785–1800.

Han, S. J., Ransom, K., Perfors, A., & Kemp, C. (2022). Human-like property induction is a challenge for large language models. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Heaven, D. (2019). Deep trouble for deep learning. *Nature*, *574*(7777), 163–166.

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Forbes, M., Borchardt, J., ... Choi, Y. (2021). *Delphi: Towards machine ethics and norms.*

Joo, S., & Yousif, S. R. (2022). Are we teleologically essentialist? *Cognitive Science*, *46*(11), e13202.

Kalish, C. W. (1995). Essentialism and graded membership in animal and artifact categories. *Memory & Cognition*, *23*(3), 335–353.

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., ... Lenci, A. (2022). Event knowledge in large language models: the gap between the impossible and the unlikely. *ArXiv*, *abs/2212.01488*.

Keil, F. C. (1992). *Concepts, kinds, and cognitive development*. mit Press.

Kelemen, D. (1999). Why are rocks pointy? children's preference for teleological explanations of th natural world. *Developmental psychology*, *35*(6), 1440.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*(1), 138–143.

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of experimental psychology: General*, *142*(4), 1074.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems.*

Kripke, S. A. (1972). Naming and necessity. In *Semantics of natural language* (pp. 253–355). Springer.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). *Package 'emmeans'*.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204.

Lombrozo, T., Kelemen, D., & Zaitchik, D. (2007). Inferring design: Evidence of a preference for teleological explanations in patients with alzheimer's disease. *Psychological Science*, *18*(11), 999–1006.

Lombrozo, T., & Rehder, B. (2012). Functions in biological kind classification. *Cognitive psychology*, *65*(4), 457–485.

Medin, D., & Ortony, A. (1989). Psychological essentialism. *Similarity and analogical reasoning*, 179–195.

Neufeld, E. (2021). Against teleological essentialism. *Cognitive Science*, *45*(4), e12961.

Newman, G. E., & Keil, F. C. (2008). Where is the essence? developmental shifts in children's beliefs about internal features. *Child development*, *79*(5), 1344–1356.

Piantadosi, S., & Hill, F. (2022). Meaning without reference in large language models. In *Neurips 2022 workshop on neuro causal and symbolic ai (ncsi).*

Putnam, H. (1962). The analytic and the synthetic.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Rose, D. (2015). Persistence through function preservation. *Synthese*, *192*(1), 97–146.

Rose, D. (2019). 14 cognitive science for the revisionary metaphysician. *Metaphysics and Cognitive Science*.

Rose, D. (2020). Mentalizing objects. *Oxford Studies in Experimental Philosophy*.

Rose, D., & Nichols, S. (2019). Teleological essentialism. *Cognitive Science*, *43*(4), e12725.

Rose, D., & Nichols, S. (2020). Teleological essentialism: generalized. *Cognitive science*, *44*(3), e12818.

Rose, D., & Schaffer, J. (2017). Folk mereology is teleological. *Experimental metaphysics*, 135–186.

Rose, D., Schaffer, J., & Tobia, K. (2020). Folk teleology drives persistence judgments. *Synthese*, *197*(12), 5491–5509.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E.-J., Ili'c, S., Hesslow, D., ... Wolf, T. (2022). Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, *abs/2211.05100*.

Talmor, A., Elazar, Y., Goldberg, Y., & Berant, J. (2019). olmpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, *8*, 743-758.

Waxman, S., Medin, D., & Ross, N. (2007). Folkbiological reasoning from a cross-cultural developmental perspective: early essentialist notions are shaped by cultural beliefs. *Developmental psychology*, *43*(2), 294.

Weir, N., Poliak, A., & Durme, B. V. (2020). Probing neural language models for human tacit assumptions. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42th annual meeting of the cognitive science society - developing a mind: Learning in humans, animals, and machines, cogsci 2020, virtual, july 29 - august 1, 2020.* cognitivesciencesociety.org.

Zhang, H., Li, L. H., Meng, T., Chang, K.-W., & den Broeck, G. V. (2022). On the paradox of learning to reason from data. *ArXiv*, *abs/2205.11502*.