UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Learning visual appearance from language is mediated by causal intuitive theories

Permalink

https://escholarship.org/uc/item/7cr4t81h

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Hauptman, Miriam
Keil, Sophia
Landau, Barbara
et al.

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Learning visual appearance from language is mediated by causal intuitive theories

Miriam Hauptman (mhauptm1@jhu.edu)
Sophia Keil (skeil4@jhu.edu)
Department of Psychological and Brain Sciences, Johns Hopkins University
Baltimore, MD, USA

Barbara Landau (landau@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
Baltimore, MD, USA

Marina Bedny (marina.bedny@jhu.edu)
Department of Psychological and Brain Sciences, Johns Hopkins University
Baltimore, MD, USA

Abstract

What and how do people learn about visual appearance from language? We test the hypothesis that in the absence of sensory evidence, people born blind use abstract causal knowledge to infer object appearance. Congenitally blind (n=19) and sighted adults (n=59) reported how many colors two types of artifacts were likely to have: artifacts for which having many colors is intended to facilitate function (n=30, e.g., fairytale book, fruit candies), and artifacts for which colorfulness is irrelevant or distracting (n=30, e.g., instructional manual, painkillers). The number of colors estimated per object was highly correlated across groups. Blind and sighted people assigned more colors to artifacts for which colorfulness facilitates function and appealed to makers' intentions in open-ended explanations. A text-only version of GPT-4 generated similar but non-identical colorfulness estimates compared to humans. Our findings suggest that people infer the appearance of unseen objects using causal 'intuitive theories' informed by linguistic evidence.

Keywords: causal reasoning; cognitive development; concepts and categories; intelligent agents; language and thought

Introduction

Language provides rich information about the sensory world. Languages of the world have large 'sensory lexicons' (e.g., 'blue,' 'round,' 'rough'; Viberg, 1983; Levinson & Majid, 2014; San Roque et al., 2015; Winter et al., 2018). Recent work with large language models (LLMs) trained exclusively on text suggests that sensory information can be acquired from language alone. For instance, LLMs can report the colors of common objects (e.g., strawberries are red), reconstruct color similarity space (e.g., red is more similar to orange than to green) and 'draw' approximate object shapes (Abdou et al., 2021; Patel & Pavlick, 2022; Sharma et al., 2024; Marjieh et al., 2022, 2024; Mukherjee et al., 2024).

Evidence from people born blind suggests that humans can also acquire knowledge about appearance from language (Marmor, 1978; Zimler & Keenan, 1983; Landau & Gleitman, 1985; Shepard & Cooper, 1992; Connolly et al., 2007; Lenci et al., 2013; Saysani et al., 2018, 2021; Kim et al., 2019, 2021; Bedny et al., 2019; Wang et al., 2020; Hauptman, Elli, et al., 2025). Landau and Gleitman (1985) showed that a 4-year-old blind child, Kelli, knew that color is a property of physical objects perceptible only with the eyes. Blind and sighted adults have shared knowledge of color similarity space, use similar color labels for some common objects (e.g., strawberries are red), and have shared intuitions about how object color varies across instances (e.g., two strawberries are more likely to be the same colors than two cars; Marmor, 1978; Shepard & Cooper, 1992; Saysani et al., 2018, 2021; Kim, et al., 2019, 2021). Although other senses, such as audition and touch, can provide information about visual experience via analogy, visual phenomena like color have no auditory or tactile analogs, pointing to language as an important source of information.

A key outstanding question is how people learn visual information from linguistic evidence. Knowledge of visual appearance can be learned in part from explicit descriptions, such as, 'Skittles are small, round, and smooth-coated candies that come in a variety of bright colors, including red, orange, and green' (GPT-4). Knowledge of one object's appearance (e.g., Skittles) can also be generalized to a broader category (e.g., fruit-flavored candy) or other related examples (e.g., M&Ms). Many modern accounts of how sensory information is learned from language emphasize tracking co-occurrence statistics of words (e.g., 'yellow' occurs with 'banana,' 'ripe' occurs with 'banana,' therefore 'yellow' and 'ripe' are related; Lewis et al., 2019; Lupyan & Lewis, 2019; Ostarek et al., 2019; Liu et al., 2025; see also Grand et al., 2022).

However, human learning often goes beyond the surface properties of the evidence. Children and adults use causal mental models, also called 'intuitive theories,' to infer unseen properties of objects, such as what objects contain on the inside based on whether they are natural kinds or artifacts (Ortony & Medin, 1989; Gelman, 2003; Keil, 1989, 1992; Bloom, 1996; Kelemen, 1999; see also Wellman & Gelman, 1992; Gopnik & Meltzoff, 1997; Tenenbaum et al., 2007; Carey, 2011; Gerstenberg & Tenenbaum, 2016). Even young children know that an object's appearance is related to deeper causal properties (Gelman & Wellman, 1991; Rosengren et al., 1991; Springer & Keil, 1991; Gopnik et al., 2001; Matan

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

& Carey, 2001; Greif et al., 2006; Gelman, 1998). For instance, children prefer biological explanations for how flowers get their colors (e.g., because of sun and rain) but intentional explanations for how cans get their colors (e.g., because someone wanted them to be that color; Springer & Keil, 1991). Such knowledge allows children to make predictions about unobservable properties from sensory evidence, e.g., if it looks like a machine, it was likely made by a person.

We hypothesized that when learning about visual appearance from language, people born blind reverse-engineer this approach: they generatively infer appearance information not accessible through sight from deeper object properties. In particular, we tested the hypothesis that for man-made objects (i.e., artifacts), people born blind use the makers' intentions to infer how colorful an object is likely to be. Consider the case of a handful of fruit candies. How many colors is it likely to have: 1, 5, or 70? Some of these answers seem more probable than others: the colors of fruit candies often communicate different flavors (of which there are typically a few). Likewise, a fairy tale book is designed to be colorful to entertain and capture the attention of children, whereas an instructional manual is less colorful so as to not distract from its content. In the current study, we tested whether people born blind would use information about an artifact's intended function to infer its colorfulness.

We presented congenitally blind and sighted adults with labels of artifacts (e.g., 'a fairy tale book') and asked them to estimate how many colors (i.e., 'colorfulness') each object was likely to have. For half of the objects, like the fairy tale book, colorfulness is intended to facilitate function, and for the other half, colorfulness is either not relevant or obtrusive to function (e.g., an instructional manual).

A separate group of participants rated each of the objects on the degree to which having many colors was useful to its function. These ratings were used to validate the object conditions and to predict participants' estimates of colorfulness item-by-item.

While sighted people could estimate the number of colors an object has based on their visual experience, people born blind cannot. We hypothesized that blind people would estimate an object's colorfulness using causal intuitions about the relationship between object colorfulness and function (e.g., colorfulness facilitates function by enhancing communicability or aesthetic appeal). If so, we predicted that i) blind and sighted people's colorfulness estimates would be correlated, ii) blind people would assign more colors to artifacts for which having many colors facilitates function, and iii) blind (and sighted) people would invoke the makers' intentions when asked to provide open-ended explanations for their colorfulness estimates.

To offer further evidence that language could serve as a source of information about colorfulness, we presented the same task to an LLM trained exclusively on text (GPT-4). Unlike blind people, LLMs do not have access to other sensory information. Thus, if GPT-4 can produce human-like judgments of colorfulness, this suggests that colorfulness information can be extracted from language. Whether GPT-4 learns something like a causal model of appearance or instead uses vast amounts of linguistic data and memory resources to mimic human-like performance remains an open question, to which we return in the Discussion.

Method

Participants

Nineteen congenitally blind adults (14 women, 5 men; age range 20-75 years, M = 40.9 ± 15 SD) and fifty-nine sighted age- and education-matched controls (37 women, 20 men, 2 non-binary people; age range: 21-71 years, M = 40.15 ± 11 SD) participated in the study. The sighted group consisted of two samples (n=19, n=40). The second sample was used as a reference group in correlation analyses. Sample size was comparable to prior studies with blind participants (e.g., Saysani et al., 2018; Kim et al., 2021) and was confirmed by a power analysis.

Blind participants lost their sight due to pathologies of the eyes or optic nerve anterior to the optic chiasm (i.e., not due to brain damage), and had at most minimal light perception since birth. All participants were screened for cognitive and neurological disabilities (self-report).

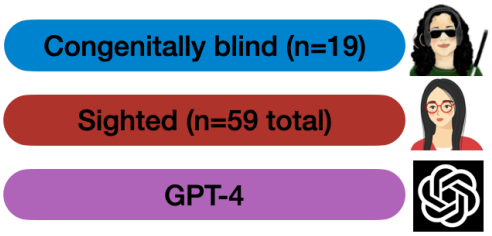
Blind participants completed the study in-person at the National Federation of the Blind National Convention and were compensated $30 per hour. Sighted participants completed the study online via Prolific and were compensated $13 per hour. Experimental procedures were reviewed and approved by the Johns Hopkins University Homewood Institutional Review Board.

Stimuli

Artifacts were organized into two conditions. Colorfulness-Intent (*Intent*) artifacts were those for which having many colors is intended to facilitate function by communicating information (e.g., candy flavor) or appealing to the aesthetic preferences of an intended audience (e.g., children) (n=30). No-Colorfulness-Intent (*No-Intent*) artifacts were those for which colorfulness is *not* intended to facilitate function; in other words, colorfulness is either irrelevant or obtrusive to function (e.g., instructional manual, mild painkillers) (n=30). Artifact descriptions were created in *Intent*/*No-Intent* pairs across conditions to control for shape. For example, fairy tale book and instructional manual formed a pair ('book'), and fruit candies and mild painkillers formed a pair ('capsules').

In an online study (n=20), a separate group of sighted participants rated how helpful having many colors is to the function of each object on a scale from 1 (not helpful) to 7 (very helpful). *Intent* objects received significantly higher 'helpful-to-function' ratings than *No-Intent* objects ($F_{(1,19)} = 24.41$, $p < .001$). These ratings were also used in correlation analyses to predict object colorfulness estimates item-by-item for both blind and sighted participants. If participants are using helpfulness-to-function to estimate number of colors, then these metrics should be correlated.

A



B

Task: How many colors do you think this object has?

object	colorfulness intent (n=30) ●	no colorfulness intent (n=30) ○
book	A fairy tale book.	An instructional manual.
towel	A towel from a beach-side shop.	A towel from a three-star hotel.
string lights	A string of Christmas lights.	A string of outdoor lights for the yard.
notebook	A diary used by a teenager in high school.	A songbook held by singers at a choir concert.
card	A get well soon card for a grandma.	A namecard for a panelist at a VIP event.
capsules	A handful of fruit candies.	A handful of mild painkillers.
paper	A park map.	A legal document.

C

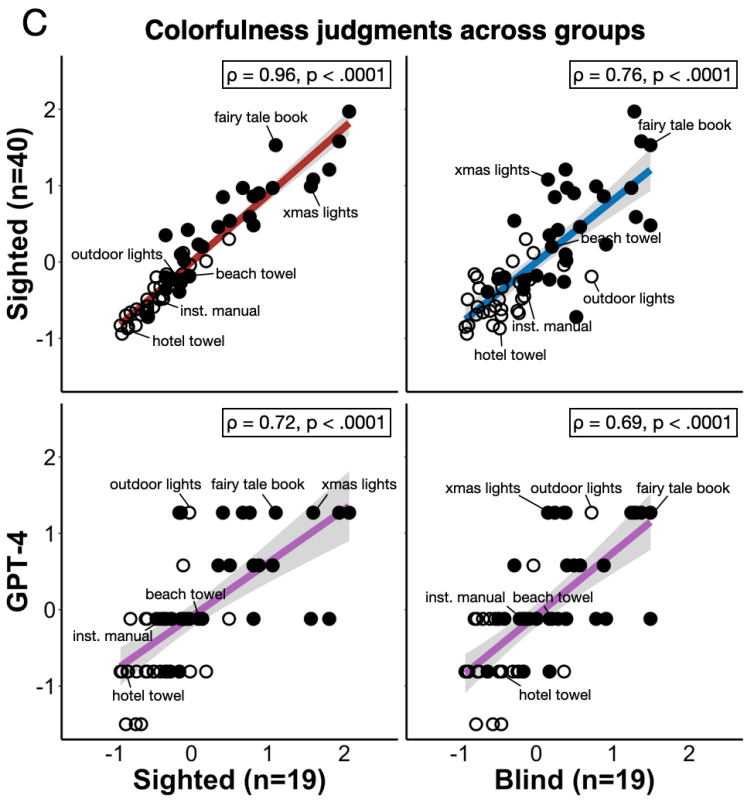


Figure 1: Shared knowledge of artifact colorfulness across groups. Panel A: Participants. The sighted group consisted of one sample of equal size to the blind group (n=19), collected first, and a larger reference sample (n=40). Panel B: Experimental design. Seven example artifact pairs (out of a total of 30 pairs) are shown for display purposes. Panel C: Item-wise correlations (Spearman's rho ρ) of normalized colorfulness judgments across groups. ρ values denote correlations between averaged group responses (e.g., average sighted vs. average blind). Confidence intervals (95%) are indicated via shading. In the plots displaying correlations with GPT-4, one outlier is excluded ('tape for scrapbooking': GPT-4: 4.75, sighted: -0.57, blind: 0.5).

Procedure

Participants provided open-ended numerical estimates of how many colors artifacts were likely to have (Figure 1A-B). The experiment had two parts. First, participants were presented with artifact labels one at a time and were asked to guess the number of colors each artifact was likely to have. All analyses of colorfulness judgments were performed using these initial judgments, which were made before the other components of the experiment were known to the participants.

In the second part of the experiment, we asked participants to guess the number of colors again, rate their confidence on a scale from 1 (not at all confident) to 7 (very confident), label the colors of the artifacts, and explain the reasoning behind their answers. Each participant was asked about 15 artifacts per condition (15 *Intent*, 15 *No-Intent*) in one of four versions, counterbalanced across participants, such that individual participants were not asked about both artifacts within a given *Intent/No-Intent* pair.

We also interrogated a text-only version of GPT-4 (gpt-4-0613, temperature=0.5) using the same task. Each artifact description was presented one at a time along with the same set of questions we asked human participants. No further information was provided.

Analyses

Colorfulness judgments were standardized (z-scored to mean=0 ± 1 SD) within each participant to account for individual differences. To assess agreement in colorfulness judgments within and across groups, we performed Spearman's rho (ρ) rank correlations. We first asked whether participant groups agreed about the degree of colorfulness of all artifacts. Colorfulness judgments made by each participant in the blind and sighted groups (both n=19) were correlated with judgments averaged across members of the sighted reference group (n=40) and judgments made by GPT-4. The significance of these correlations was then tested using Student's t-tests and ANOVAs that modeled the Fisher-Z transformed single-subject correlations. To assess agreement within the blind group, we calculated Kendall's coefficient of concordance (*W*) to account for the fact that there was no blind reference group.

Next, we asked whether participant groups agreed about differences in the number of colors assigned to artifacts within *Intent*/*No-Intent* pairs (e.g., 'fairy tale book' minus

'instructional manual'). We correlated average artifact pair differences for each group.

Finally, to assess agreement in the use of color labels across groups, we selected the most frequent color labels (n=13 color labels; accounting for 97% of all labels) used by blind and sighted participants combined. We then correlated across groups the relative differences in the use of these labels for *Intent* vs. *No-Intent* artifacts.

Results

Blind and sighted people agree about how many colors objects have

The number of colors generated per object (i.e., colorfulness judgments) were highly correlated across the two sighted samples, suggesting shared intuitions among sighted people (sighted vs. sighted reference, average of individual participants' correlation values: $\rho = .72 \pm 0.15$ SD; one sample t-test: $t_{(18)} = 13.65$, $p < .001$; Figure 1C). Colorfulness judgments of blind and sighted people were also significantly correlated (average of individual participants' correlation values: $\rho = .59 \pm 0.14$ SD; $t_{(18)} = 13.76$, $p < .001$), though less so than between the two sighted samples (two-sample t-test, $t_{(36)} = -3.02$, $p = .005$). Blind participants also showed high within-group agreement (blind: Kendall's $W = .54$, $p < .001$, sighted: $W = .57$, $p < .001$). These results suggest that first-person sensory access is not required to estimate how many colors an object has.

Blind and sighted people infer object colorfulness from intended function

Both sighted and congenitally blind participants judged Colorfulness-Intent (*Intent*) artifacts to have more colors than No-Colorfulness-Intent (*No-Intent*) artifacts (subject-wise repeated measures ANOVA, 2 conditions (*Intent*, *No-Intent*) x 2 groups (sighted, blind): main effect of condition, $F_{(1,36)} = 441.29$, $p < .001$; Figure 2).

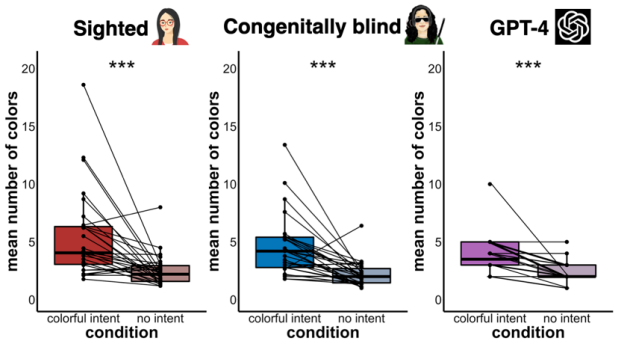


Figure 2: Average item-wise colorfulness judgments across groups. Average non-normalized judgments for all 60 artifacts are displayed for each group. Paired artifacts (e.g., 'fairy tale book,' 'instructional manual') are connected by lines.

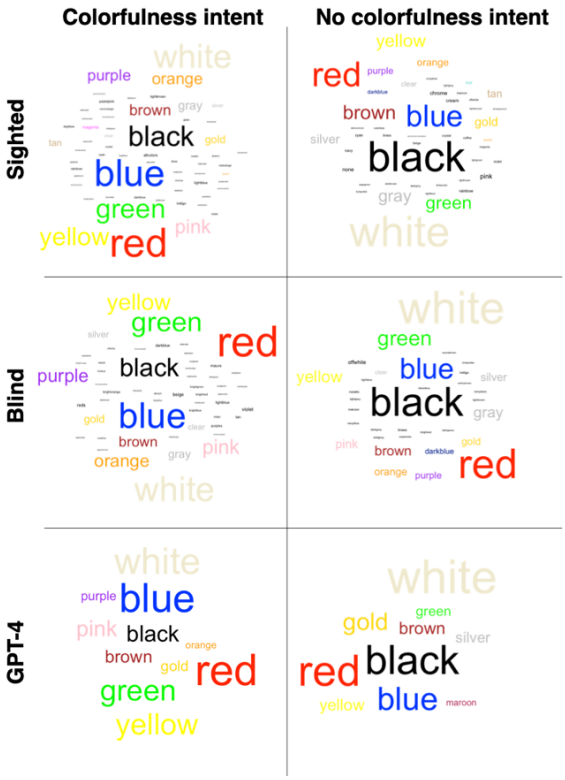


Figure 3: Color label use across groups and conditions. The size of each color label denotes its relative frequency within participant groups and conditions.

All groups assigned more colors to artifacts for which colorfulness is intended to facilitate function (i.e., Colorfulness-Intent artifacts; all *p*s < .001). This effect was highly significant in each group and statistically identical across groups (group by condition interaction, $F_{(1,36)} = 0.03$, $p = .88$; main effect of group, $F_{(1,36)} = 1$, $p = .32$).

In both groups, colorfulness judgments were highly and equally correlated with ratings of how helpful colorfulness is to function collected in a separate online study (sighted: $\rho = .84$, $p < .001$, blind: $\rho = .82$, $p < .001$). These results support the hypothesis that people born blind use this information to predict the number of colors an object is likely to have.

Further evidence for the idea that blind people can infer colorfulness from intention comes from analyses of item-wise variation. Since items within a pair were matched on shape, differences in the number of colors assigned within *Intent*/*No-Intent* pairs offers a more fine-grained test of whether the degree to which colorfulness facilitates function influences estimates of colorfulness. Difference scores were highly correlated across the sighted samples ($\rho = .86$, $p < .001$) and were also correlated across sighted and congenitally blind people ($\rho = .51$, $p = .004$). Notably, the two sighted samples exhibited numerically higher agreement than sighted and blind participants, suggesting that sighted people may rely partially on visual memory when making

colorfulness judgments, whereas people born blind may rely more on inferences about intention.

Finally, we observed numerical differences in the color labels that blind and sighted people used across *Intent* and *No-Intent* artifacts, whereby brighter colors (e.g., blue, red, green) were produced more often for *Intent* artifacts and neutral colors (e.g., black, white) were produced more often for *No-Intent* artifacts (Figure 3). Differences in color label use across conditions was highly consistent across groups ($\rho = .98$, all ps $< .001$).

Together, these results suggest that people born blind infer colorfulness of artifacts by appealing to the intentions of the maker (e.g., fairy tale books are colorful because they are intended to capture the reader's attention; Table 1).

Table 1: Example explanations from each group. For each object, participants were asked, "why did you choose that number of colors?"

	Fairy tale book	Instructional manual
Sighted	The ones we own are colorful.	They are usually not too colorful from my experience.
Blind	They are usually geared toward children, and colors grab a kid's attention.	It would have to have one color to be easy to find among other booklets.
GPT-4	Most have colorful illustrations to make the stories more engaging.	Most I've seen usually have a base color, a color for accents, and black text.

Humans and GPT-4 generate similar but non-identical colorfulness judgments

To provide additional insight into the extent to which colorfulness information is available in linguistic data, we interrogated a text-only version of GPT-4 using the same task. Unlike blind people, GPT-4 lacks access to sensory information, including from touch, audition, and smell. If GPT-4 nevertheless produces similar colorfulness estimates to humans, this would suggest that language contains relevant information about object colorfulness.

Colorfulness judgments were significantly correlated across human participants and GPT-4, an LLM trained exclusively on text (blind vs. GPT-4, average of individual blind participants' correlation values: $\rho = .5 \pm 0.16$ SD; sighted vs. GPT-4: $\rho = .5 \pm 0.12$ SD, both ps $< .001$). Correlations between human groups (blind vs. sighted reference, sighted vs. sighted reference) were higher than correlations between humans and GPT-4 (main effect of comparison group (sighted reference, GPT-4), $F_{(1,36)} = 54.45$, $p < .001$). This finding suggests that although GPT-4 can acquire information about artifact colorfulness from language, it does not fully capture human performance.

Like blind and sighted people, GPT-4 assigned more colors to artifacts for which colorfulness is intended to facilitate function. (Figure 2; *Intent* vs *No-Intent* artifacts $t_{(58)} = 4.56$, $p < .001$). The size of this effect was not different between GPT-4, and the human groups (item-wise repeated measures ANOVA, 2 conditions (*Intent*, *No-Intent*) x 3 groups (sighted, blind, GPT-4), group by condition interaction, $F_{(2,116)} = 0.27$, $p = .76$; main effect of group, $F_{(2,116)} = 0$, $p = 1$).

Differences in the number of colors within *Intent*/*No-Intent* pairs were also correlated across GPT-4 and congenitally blind participants (GPT-4 vs. congenitally blind: $\rho = .6$, $p < .001$). Interestingly, colorfulness difference scores were not correlated across GPT-4 and sighted participants (GPT-4 vs. sighted: $\rho = .23$, $p = .21$). However, as reported above, blind and sighted people's difference scores were correlated ($\rho = .51$, $p = .004$). This pattern of results suggests that blind people and GPT-4 learn object colorfulness from language in partially different ways.

In sum, our findings from a text-only version of GPT-4 suggest that linguistic evidence contains information about the number of colors objects have, but that humans and large language models do not learn this information the same way.

Discussion

We find that blind and sighted people living within the same cultural context show high agreement about how many colors an artifact is likely to have. Sighted adults likely base their colorfulness judgments in part on common visual experiences. Consistent with this idea, they agree more with each other than with congenitally blind people or a text-trained LLM. Sighted participants also appeal to their experiences with real objects when explaining their colorfulness judgments (e.g., 'I said that fairy tale books have nine colors because the fairy tale book I own has that many colors'). By contrast, people born blind acquire their knowledge of object colorfulness from linguistic evidence.

How do people and machines learn how many colors an object is likely to have from language? As noted in the Introduction, some 'visual' information can be learned from explicit descriptions. For example, most blind people agree that school buses are yellow, which can only be learned as an explicitly stated fact (Kim et al., 2021). However, learning appearance only from explicit descriptions or even by generalizing from explicit descriptions of one object to another has limitations. Stated color facts are only useful for previously encountered objects, or objects that are similar to those objects.

Relying exclusively on such fact-based learning is feasible for LLMs, which are trained on vast amounts of linguistic data and have better memory capacities but is less reliable for humans. Blind and sighted people often disagree on arbitrary object-color mappings (e.g., in one study, 50% of blind

participants and 100% of sighted participants labeled bananas yellow and polar bears white; Kim et al., 2019; 2021). By contrast, modern LLMs readily generate the colors of objects that align with the intuitions of sighted people (e.g., Liu et al., 2025). For example, text-trained GPT-4 produces 'brown' for aardvarks and 'pink' for axolotls. However, LLMs that are more modest in size also sometimes do not acquire sighted-like object-color pairings, and, unlike blind people, also appear to be fooled by raw frequency statistics (i.e., uncommon colors are named more frequently, such as 'green' for banana; Liu et al., 2025).

The current evidence suggests that in addition to using learning mechanisms that rely on explicitly stated descriptive facts, humans also rely on a generative approach to infer unobservable perceptual properties of objects using intuitive theories of why objects appear the way they do. In particular, we find that blind adults provide higher colorfulness estimates for objects for which colorfulness is intended to facilitate function and also appeal to the intentions of the makers in their explanations.

The current results are consistent with prior evidence that blind and sighted people tend to agree about visual knowledge that can be inferred from deeper causal properties (Kim et al., 2019; 2021). For example, blind and sighted people have similar intuitions about how object color varies across tokens (e.g., two stop signs are more likely to be the same color than two cars), an aspect of appearance that can be explained by the causal relationship between object function and appearance (i.e., object color is less likely to vary when it is relevant to function; Kim et al., 2021). Here we suggest that an analogous mechanism is used to infer the colorfulness of individual object tokens. In addition to being less mnemonically costly, a generative approach also enables inferences about objects that are very different from those previously encountered in the evidence, including novel objects.

A key open question is *how* linguistic evidence is used to construct causal models of appearance, including models of how color is used to entertain, impress, distract, inform, and capture the attention of others. Another open question is whether LLMs acquire something like causal models of object colorfulness and, like humans, use such models to infer appearance. Because LLMs like GPT-4 have memory capacities and access to more data compared to human learners, it is also highly plausible that they mimic human inference using more 'memory-based' approaches (Bender & Koller, 2020; Warstadt & Bowman, 2022; Frank, 2023; Kauf et al., 2023; Lake & Murphy, 2023; McCoy et al., 2024). Although GPT-4's colorfulness judgments were significantly correlated with those of humans in the current study, its judgments for some items (e.g., scrapbooking tape) were distinctly not human-like. Endowing LLMs with causal models of appearance as a form of inductive biases during training may enable the creation of AI models that bridge language and vision in human-like ways.

In summary, the current study demonstrates that knowledge of visual appearance, specifically artifact colorfulness, can be learned from language alone. However, our findings also suggest that for humans, learning from language is mediated by causal mental models of the world, i.e., 'intuitive theories.' These intuitive theories are themselves shaped by linguistic evidence and enable generative inferences about sensory information in the absence of first-person experience.

Acknowledgments

We would like to thank all of the blind and sighted participants, the blind community, and the National Federation of the Blind. Without their support, this study would not be possible. We thank Kartik Chandra for providing thoughtful feedback on an earlier draft. This work was supported by the National Institutes of Health (R01 EY027352 to M.B.).

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021). Can language models encode perceptual Structure without grounding? A case study in color. arXiv:2109.06129. *arXiv.* https://doi.org/10.48550/arXiv.2109.06129

Bedny, M., Koster-Hale, J., Elli, G., Yazzolino, L., & Saxe, R. (2019). There's more to "sparkle" than meets the eye: Knowledge of vision and light verbs among congenitally blind and sighted individuals. *Cognition, 189,* 105–115. https://doi.org/10.1016/j.cognition.2019.03.017

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition, 60*(1), 1-29.

Carey, S. (2011). *The origin of concepts*. Oxford University Press.

Connolly, A. C., Gleitman, L. R., & Thompson-Schill, S. L. (2007). Effect of congenital blindness on the semantic representation of some everyday concepts. *Proceedings of the National Academy of Sciences, 104*(20), 8241–8246. https://doi.org/10.1073/pnas.0702812104

Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology, 2*(8), 451–452. https://doi.org/10.1038/s44159-023-00211-x

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive psychology, 20*(1), 65-95.

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

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious.

Cognition, 38(3), 213–244. https://doi.org/10.1016/0010-0277(91)90007-Q

Gerstenberg, T., & Tenenbaum, J. B. (2016). Intuitive theories. In M. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning.* Oxford University Press.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* The MIT Press.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language, 7*(1-2), 145–171.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*(5), 620–629.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, *6*(7), 975-987.

Greif, M. L., Kemler Nelson, D. G., Keil, F. C., & Gutierrez, F. (2006). What do children want to know about animals and artifacts? Domain-specific requests for information. *Psychological Science, 17*(6), 455-459.

Hauptman, M., Elli, G., Pant, R., & Bedny, M. (2025). Neural specialization for 'visual' concepts emerges in the absence of vision. *Cognition, 257,* 106058. https://doi.org/10.1016/j.cognition.2024.106058

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., Fedorenko, E., & Lenci, A. (2023). Event knowledge in Large Language Models: The gap between the impossible and the unlikely. *Cognitive Science, 47*(11), e13386. https://doi.org/10.1111/cogs.13386

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

Keil, F.C. (1989*). Concepts, kinds and cognitive development.* Cambridge, MA: Bradford Books.

Keil, F. C. (1992). The origins of an autonomous biology. In M. R. Gunnar, M. Maratsos (Eds.), *Modularity and constraints in language and cognition.* Psychology Press.

Kim, J. S., Aheimer, B., Montané Manrara, V., & Bedny, M. (2021). Shared understanding of color among sighted and blind adults. *Proceedings of the National Academy of Sciences, 118*(33), e2020192118. https://doi.org/10.1073/pnas.2020192118

Kim, J. S., Elli, G. V., & Bedny, M. (2019). Knowledge of animal appearance among sighted and blind adults. *Proceedings of the National Academy of Sciences, 116*(23), 11213–11222. https://doi.org/10.1073/pnas.1900952116

Lake, B. M., & Murphy, G. L. (2023). Word meaning in minds and machines. *Psychological Review, 130*(2), 401–431. https://doi.org/10.1037/rev0000297

Landau, B., & Gleitman, L. R. (1985). *Language and experience: Evidence from the blind child.* Harvard University Press.

Lenci, A., Baroni, M., Cazzolli, G., & Marotta, G. (2013). BLIND: A set of semantic feature norms from the congenitally blind. *Behavior Research Methods, 45*(4), 1218–1233. https://doi.org/10.3758/s13428-013-0323-4

Levinson, S. C., & Majid, A. (2014). Differential ineffability and the senses. *Mind & Language, 29*(4), 407–427. https://doi.org/10.1111/mila.12057

Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, *116*(39), 19237-19238.

Liu, Q., van Paridon, J., & Lupyan, G. (2025). Learning about color from language. *Communications Psychology*, *3*(1), 60.

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319-1337.

Marjieh, R., van Rijn, P., Sucholutsky, I., Sumers, T. R., Lee, H., Griffiths, T. L., & Jacoby, N. (2022). Words are all you need? capturing human sensory similarity with textual descriptors. arXiv:2206.04105. *arXiv.*

Marjieh, R., Sucholutsky, I., Van Rijn, P., Jacoby, N., & Griffiths, T. L. (2024). Large language models predict human sensory judgments across six modalities. *Scientific Reports, 14*(1), 21445. https://doi.org/10.1038/s41598-024-72071-1

Marmor, G. S. (1978). Age at onset of blindness and the development of the semantics of color names. *Journal of Experimental Child Psychology, 25*(2), 267–278. https://doi.org/10.1016/0022-0965(78)90082-6

Matan, A., & Carey, S. (2001). Developmental changes within the core of artifact concepts. *Cognition, 78*(1), 1–26. https://doi.org/10.1016/S0010-0277(00)00094-9

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences, 121*(41), e2322420121. https://doi.org/10.1073/pnas.2322420121

Medin, D., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. New York: Cambridge University Press.

Mukherjee, K., Rogers, T. T., & Schloss, K. B. (2024). Large Language Models estimate fine-grained human color-concept associations. arXiv:2406.17781. *arXiv*. https://doi.org/10.48550/arXiv.2406.17781

Ostarek, M., Van Paridon, J., & Montero-Melis, G. (2019). Sighted people's language is not helpful for blind individuals' acquisition of typical animal colors. *Proceedings of the National Academy of Sciences, 116*(44), 21972-21973.

Patel, R., & Pavlick, E. (2022). Mapping language models to grounded conceptual spaces. In *International conference on learning representations.*

Rosengren, K. S., Gelman, S. A., Kalish, C. W., & McCormick, M. (1991). As time goes by: Children's early understanding of growth in animals. *Child Development, 62*(6), 1302–1320. https://doi.org/10.1111/j.1467-8624.1991.tb01607.x

San Roque, L., Kendrick, K. H., Norcliffe, E., Brown, P., Defina, R., Dingemanse, M., Dirksmeyer, T., Enfield, N., Floyd, S., Hammond, J., Rossi, G., Tufvesson, S., Van Putten, S., & Majid, A. (2015). Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies. *Cognitive Linguistics, 26*(1), 31–60. https://doi.org/10.1515/cog-2014-0089

Saysani, A., Corballis, M. C., & Corballis, P. M. (2018). Colour envisioned: Concepts of colour in the blind and sighted. *Visual Cognition, 26*(5), 382–392. https://doi.org/10.1080/13506285.2018.1465148

Saysani, A., Corballis, M. C., & Corballis, P. M. (2021). Seeing colour through language: Colour knowledge in the blind and sighted. *Visual Cognition, 29*(1), 63–71. https://doi.org/10.1080/13506285.2020.1866726

Sharma, P., Shaham, T. R., Baradad, M., Rodriíuez-Muñoz, A., Duggal, S., Isola, P., Torralba, A., & Fu, S. (2024). A Vision Check-up for Language Models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14410–14419. https://doi.org/10.1109/CVPR52733.2024.01366

Shepard, R. N., & Cooper, L. A. (1992). Representation of colors in the blind, color-blind, and normally sighted. *Psychological Science, 3*(2), 97–104. https://doi.org/10.1111/j.1467-9280.1992.tb00006.x

Springer, K., & Keil, F. C. (1991). Early differentiation of causal mechanisms appropriate to biological and nonbiological kinds. *Child Development, 62*(4), 767–781. https://doi.org/10.1111/j.1467-8624.1991.tb01568.x

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal Learning*. Oxford University Press: New York. https://doi.org/10.1093/acprof:oso/9780195176803.003.0020

Viberg, Å. (1983). The verbs of perception: A typological study. *Linguistics 21*, 123-162.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language* (pp. 17-60). CRC Press.

Wang, X., Men, W., Gao, J., Caramazza, A., & Bi, Y. (2020). Two forms of knowledge representations in the human brain. *Neuron, 107*(2), 383-393.e5. https://doi.org/10.1016/j.neuron.2020.04.010

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: foundational theories of core domains. *Annual Review of Psychology, 43*(1), 337–375. https://doi.org/10.1146/annurev.ps.43.020192.002005

Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition, 179,* 213–220. https://doi.org/10.1016/j.cognition.2018.05.008

Zimler, J., & Keenan, J. M. (1983). Imagery in the congenitally blind: How visual are visual images? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(2), 269–282. https://doi.org/10.1037/0278-7393.9.2.269