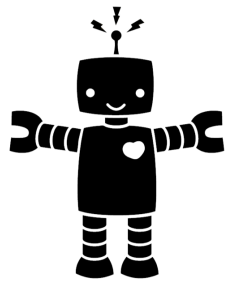




Berkeley
UNIVERSITY OF CALIFORNIA



Kidd Lab

How AI distorts human beliefs

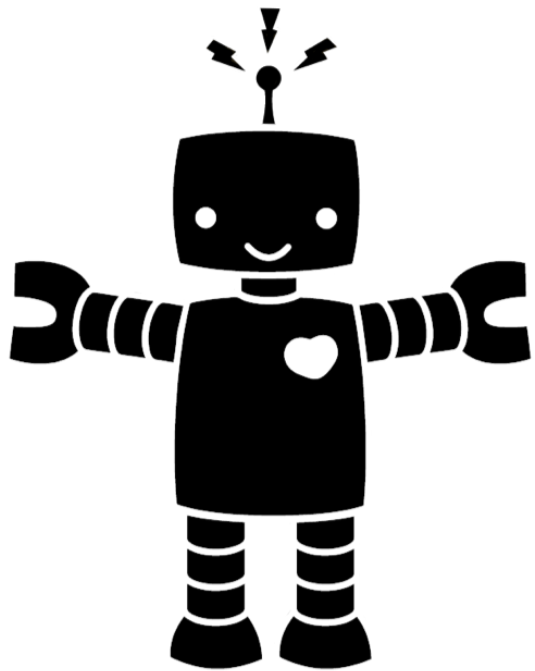
Celeste Kidd (celestekidd@berkeley.edu)

Psychology, University of California, Berkeley



@celestekidd.bsky.social

World Summit AI, Doha, Qatar, 10 Dec 2024



Kidd Lab

www.kiddlab.com



How do biologically intelligent agents form their beliefs?

How do biologically intelligent agents form their beliefs?

How do new technologies impact those beliefs?



AI And The Democratization Of Knowledge



Mark Pittman Forbes Councils Member

Forbes Technology Council COUNCIL POST | Membership (Fee-Based)



GETTY

Jun 25, 2024, 08:00am EDT

Mark Pittman is Founder of [Blyncsy, Inc.](#) & Director of Transportation AI at Bentley.

The rise of artificial intelligence (AI) is not just transforming industries; it's reshaping the very fabric of knowledge in our world. As AI continues to advance, we are witnessing a remarkable phenomenon: the flattening of the knowledge curve.

Brave New Words

How AI Will
Re|volutionize
Education (and
Why That's a
Good Thing) ☀️

Salman Khan

Founder of Khan Academy

"A timely master class for anyone interested in the future
of learning in the AI era." —Bill Gates

Today, I want to
focus on what
you're not
hearing.

A young child with dark hair is looking intently at a tablet screen. The child's face is partially visible, showing their eyes and nose. The background is dark and out of focus. The text is overlaid on the image in a large, white, sans-serif font.


Careless integration
of AI into people's
lives can compromise
their access to truth.

RECENT PROJECTS



Humans Are Biased. Generative AI Is Even Worse

Stable Diffusion's text-to-image model amplifies stereotypes about race and gender — here's why that matters.

 Built with **Svelte** and **D3.js** for **BLOOMBERG**

 December 2022

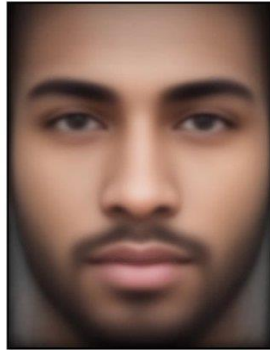
Stable Diffusion Perpetuates Criminal Stereotypes

Composite average of all images

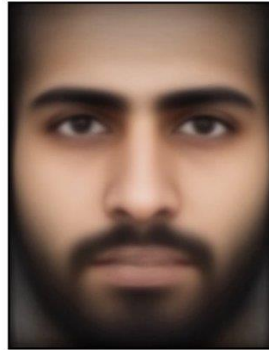
INMATE



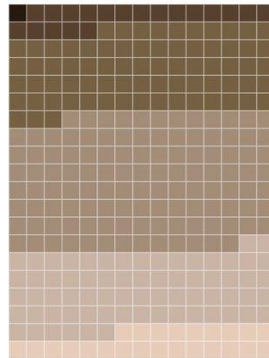
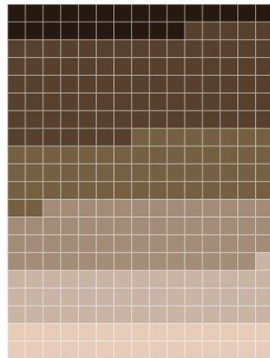
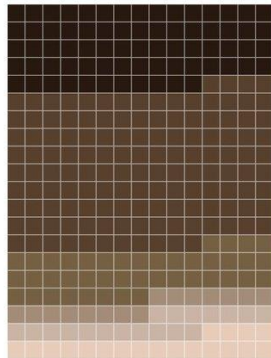
DRUG DEALER



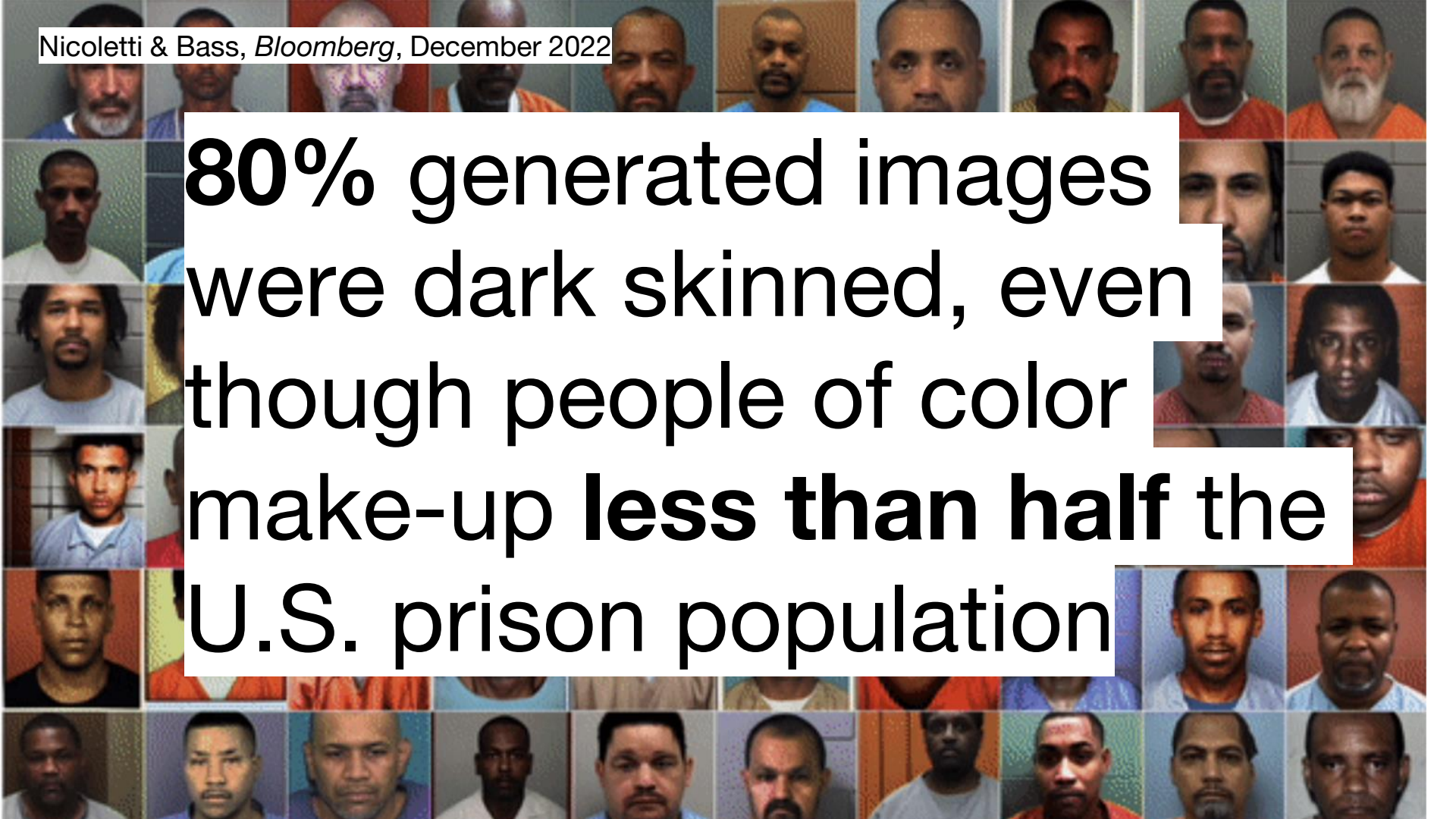
TERRORIST



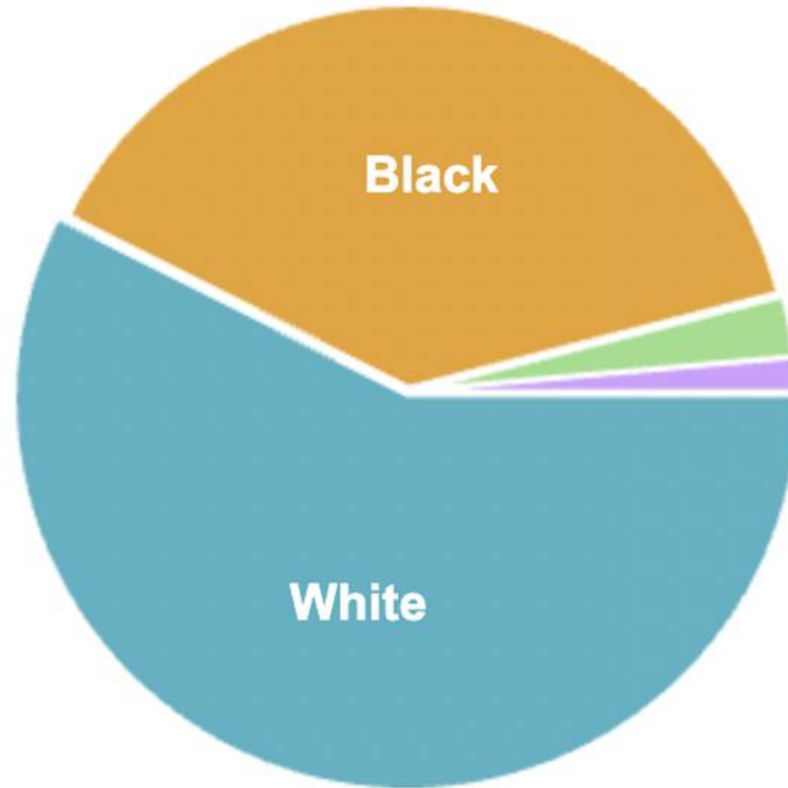
Distribution of skin tones



80% generated images were dark skinned, even though people of color make-up **less than half** the U.S. prison population



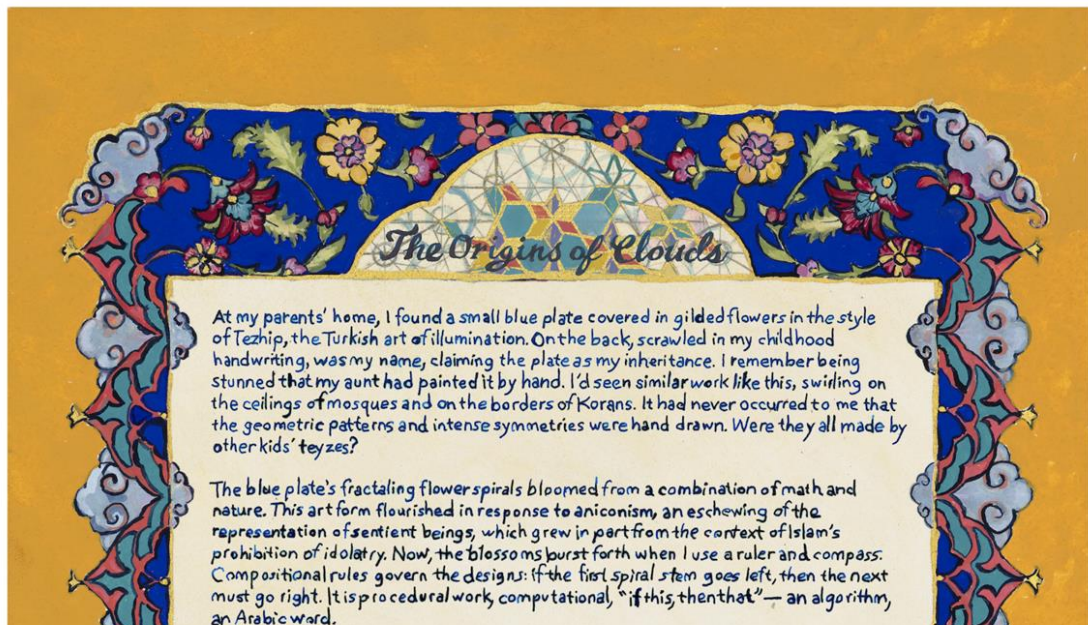
Federal Bureau of Prisons, Inmate Statistics, September 2023



Statistics are updated weekly. Last updated on Saturday, 30 September 2023

The Origin of Clouds

Şerife Wong



Şerife Wong
Icarus Salon

TECHNOLOGY

AI's Present Matters More Than Its Imagined Future

Let's not spend too much time daydreaming.

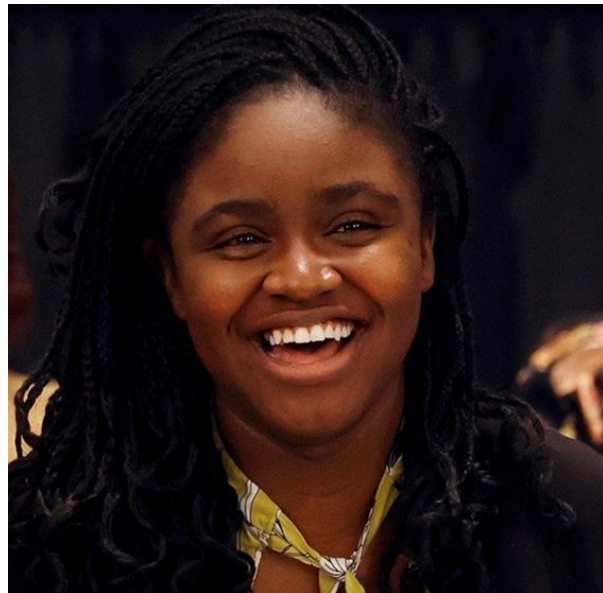
By Inioluwa Deborah Raji

OCTOBER 4, 2023

SHARE ▼ SAVED STORIES ↗ SAVE □

Last month, I found myself in a particular seat. A few places to my left was Elon Musk. Down the table to my right sat Bill Gates. Across the room sat Satya Nadella, Microsoft's CEO, and not too far to his left was Eric Schmidt, the former CEO of Google. At the other end of the table sat Sam Altman, the head of OpenAI, the company responsible for ChatGPT.

We had all arrived that morning for the inaugural meeting of Senate Leader Chuck Schumer's AI Insight Forum—the first of a set of events with an ambitious objective: to accelerate a bipartisan path toward meaningful



Deborah Raji
UC Berkeley

Gebru, *New York Times*, December 2024

The New York Times



TURNING POINTS: GUEST ESSAY

Who Is Tech Really For?

As Silicon Valley chases military tech and funding, it's losing sight of what inspires its workers.

By Timnit Gebru

Timnit Gebru is the executive director of the Distributed Artificial Intelligence Research Institute.

Dec. 5, 2024



Timnit Gebru
*Distributed Artificial
Intelligence Research
Institute (DAIR)*

Large language models and the perils of their hallucinations

[Razvan Azamfirei](#) , [Sapna R. Kudchadkar](#) & [James Fackler](#)

Salvagno et al. present a ChatGPT-generated summary of three studies. As they noted, the summary was believable, albeit generic and sparse in the details. The glaring problem is that it's completely fabricated. ChatGPT cannot access the internet, and its training dataset stops in September 2021; it has no reference to any studies published in 2023 [2]. In fact, one of the trials included in the summary, Belohlavek et al. [3], showed no improvement in functional neurological outcomes, contradicting ChatGPT's summary.

We must understand one particular aspect of large language models, which is gracefully termed as “hallucinations”, though “fabricating information” may be more accurate [4]. In the

case of the ChatGPT summary we see an embedding of a generic summary of an average study



Sam Altman ✓

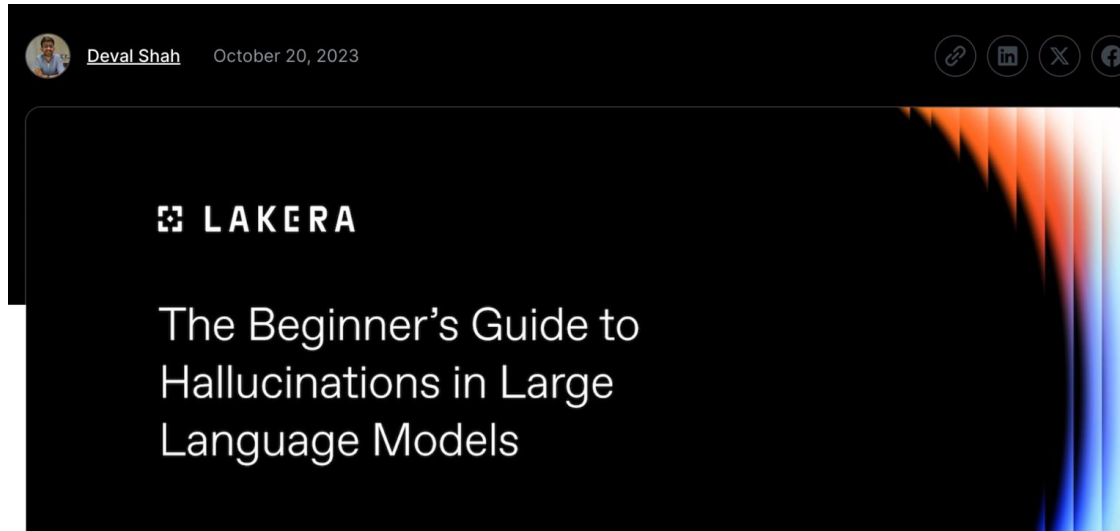
@sama



this is going to take continual iteration--and lots and lots of societal input--to get right.

to find the right balance, we will likely overcorrect several times, and find new edges in the technology. we appreciate the patience and good faith as we get to a better place!

10:17 AM · Feb 16, 2023 · **138K** Views



On this page

[Understanding LLM Hallucinations](#)

[Causes of Hallucinations in LLMs](#)

[Implications of Hallucinations](#)

[Mitigating Hallucinations in Large Language Models](#)

[Case Studies and Industry Insights](#)

[Additional Resources](#)

[Key Takeaways](#)

Large Language Models (LLMs) are at the forefront of technological discussions, known for their proficiency in processing and generating text that resembles human communication. They are transforming our interactions with technology. However, these models are not without their flaws. One significant issue is their tendency to produce "hallucinations," which affect their reliability.

Hallucinations in LLMs refer to the generation of content that is irrelevant, made-up, or inconsistent with the input data. This problem leads to incorrect information, challenging the trust placed in these models. Hallucinations are a critical obstacle in the development of

2 problematic
ways AI distorts
human beliefs

1.

1. AI transmits harmful, stubborn biases and fabrications to human users.

PSYCHOLOGY

How AI can distort human beliefs

Models can convey biases and false information to users

By **Celeste Kidd**¹ and **Abeba Birhane**^{2,3}

Individual humans form their beliefs by sampling a small subset of the available data in the world. Once those beliefs are formed with high certainty, they can become stubborn to revise. Fabrication and bias in generative artificial intelligence (AI) models are established phenomena that can occur as part of regular system use, in the absence of any malevolent forces seeking to push bias or disinformation. However, such transmission of false information and bias

communication, and the other fields that are considering the impact of bias and misinformation on population-level beliefs.

People form stronger, longer-lasting beliefs when they receive information from agents that they judge to be confident and knowledgeable, starting in early childhood. For example, children learned better when they learned from an agent who asserted their knowledgeability in the domain as compared with one who did not (5). That very young children track agents' knowledgeability and use it to inform their beliefs and exploratory

BRIEF REPORT



The role of prior knowledge and curiosity in learning

Shirlene Wade^{1,2} · Celeste Kidd¹

Published online: 11 May 2019
© The Psychonomic Society, Inc. 2019

Abstract

Recent work has argued that curiosity can improve learning. However, these studies also leave open the possibility that being on the verge of knowing can itself induce curiosity. We investigate how prior knowledge relates to curiosity and subsequent learning using a trivia question task. Curiosity in our task is best predicted by a learner's estimate of their current knowledge, more so than an objective measure of what they actually know. Learning is best predicted by both curiosity and an objective measure of knowledge. These results suggest that while curiosity is correlated with knowledge, there is only a small boost in learning from being curious. The implication is that the mechanisms that drive curiosity are not identical to those that drive learning outcomes.

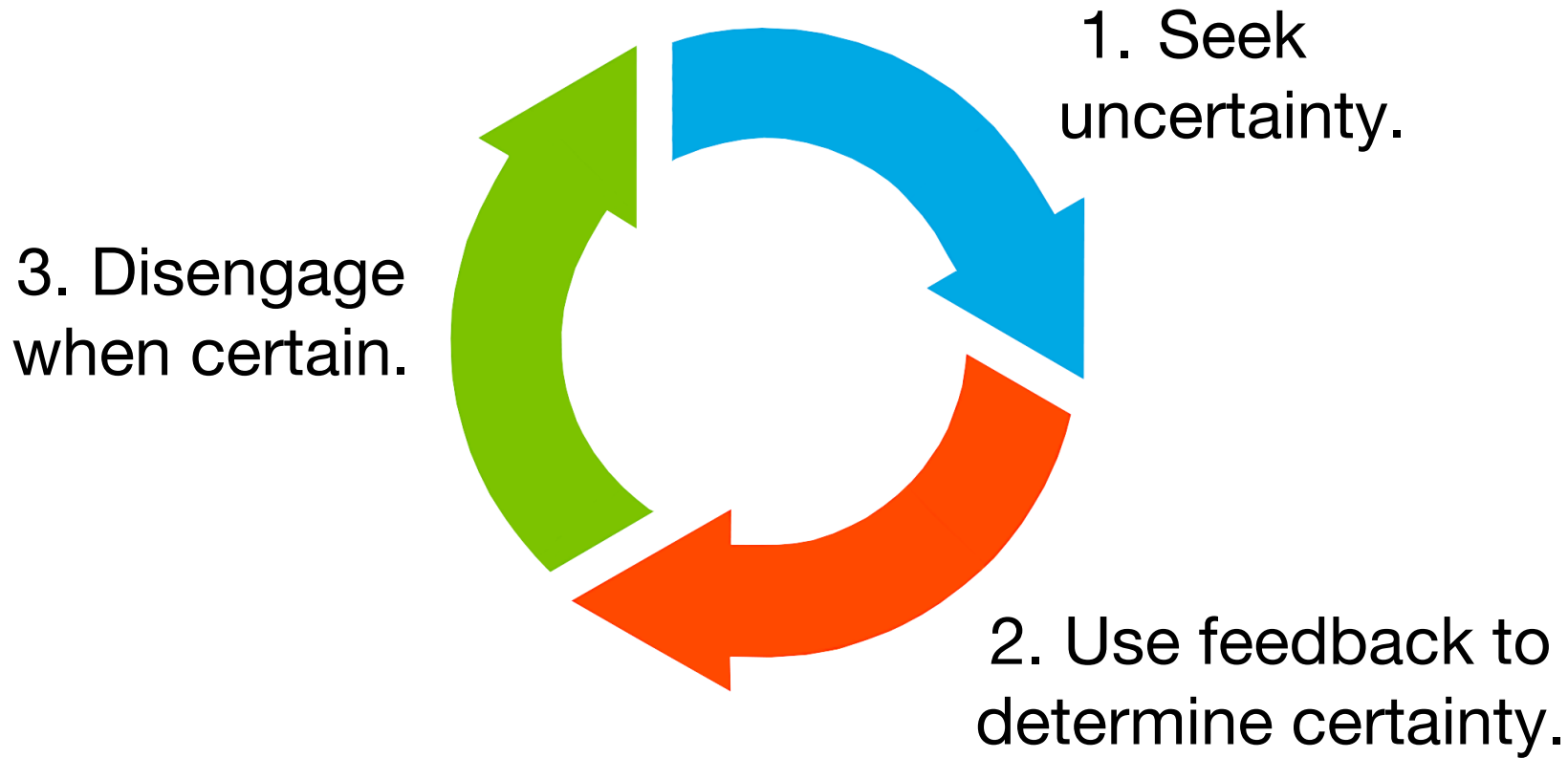
Keywords Curiosity · Memory · Learning · Metacognition

Introduction

inferior frontal gyrus – regions related to long-term memory consolidation – were modulated by the individual's level of cu-



Curiosity cycle



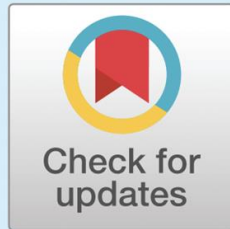
Mari, Molilica, Piantadosi, & Kidd, *Open Mind*, 2018

Yang, Martí, Baer, Granera, Palmeri, & Kidd, *CogSci*, 2024



Discoveries in
Cognitive Science

an open access  journal



Citation: Martí, L., Mollica, F.,

Certainty Is Primarily Determined by Past Performance During Concept Learning

Louis Martí^{1,2}, Francis Mollica¹, Steven Piantadosi^{1,2}, and Celeste Kidd^{1,2}

¹Brain and Cognitive Sciences, University of Rochester, Rochester

²Psychology, University of California, Berkeley

Keywords: certainty, confidence, metacognition, learning, concepts

ABSTRACT

Prior research has yielded mixed findings on whether learners' certainty reflects veridical probabilities from observed evidence. We compared predictions from an idealized model of learning to humans' subjective reports of certainty during a Boolean concept-learning task in order to examine subjective certainty over the course of abstract, logical concept learning. Our analysis evaluated theoretically motivated potential predictors of certainty to determine how well each predicted participants' subjective reports of certainty. Regression analyses that controlled for individual differences demonstrated that despite learning curves tracking the ideal learning models, reported certainty was best explained by performance rather than measures derived from a learning model. In particular, participants' confidence was driven primarily by how well they observed themselves doing, not by idealized statistical inferences made from the data they observed.

INTRODUCTION

Daily life requires making judgments about the world based on inconclusive evidence. These



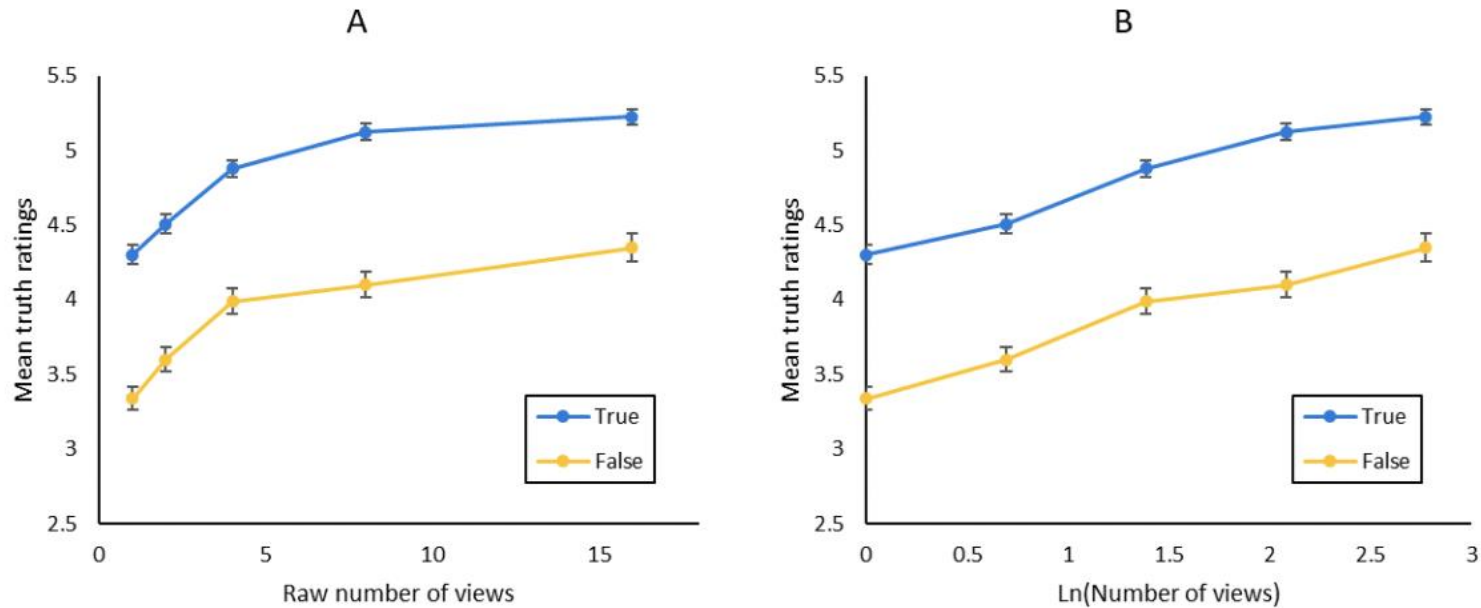
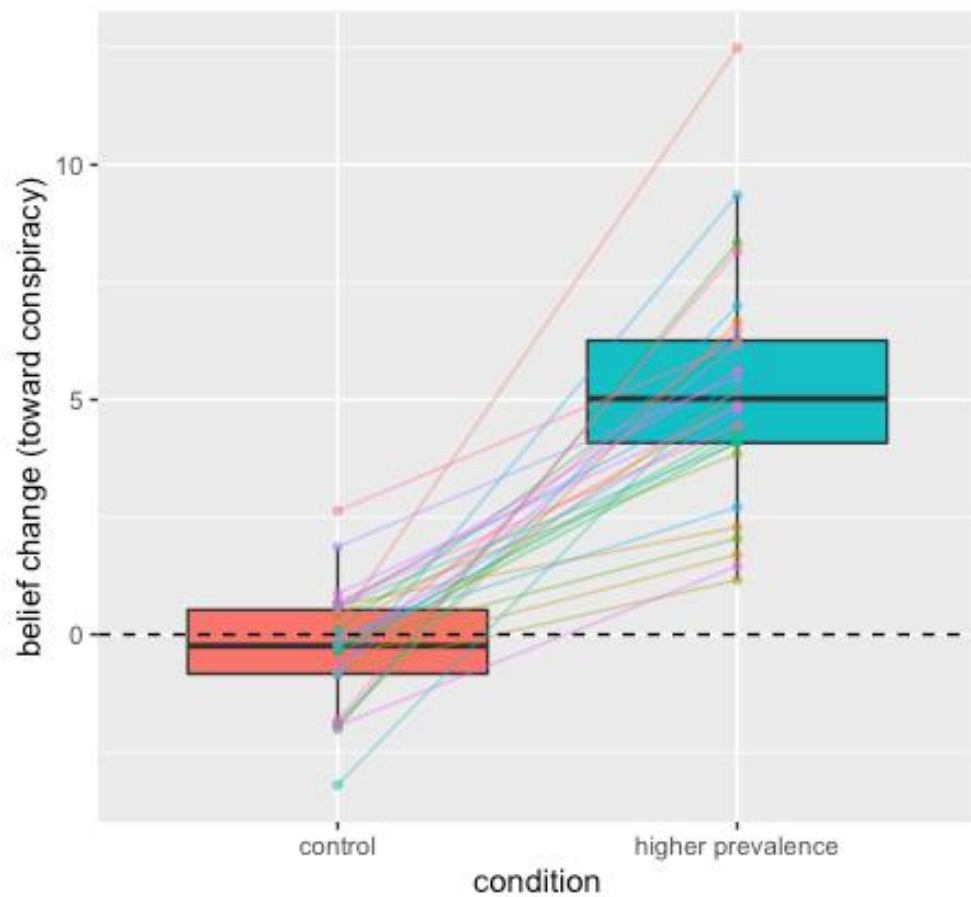
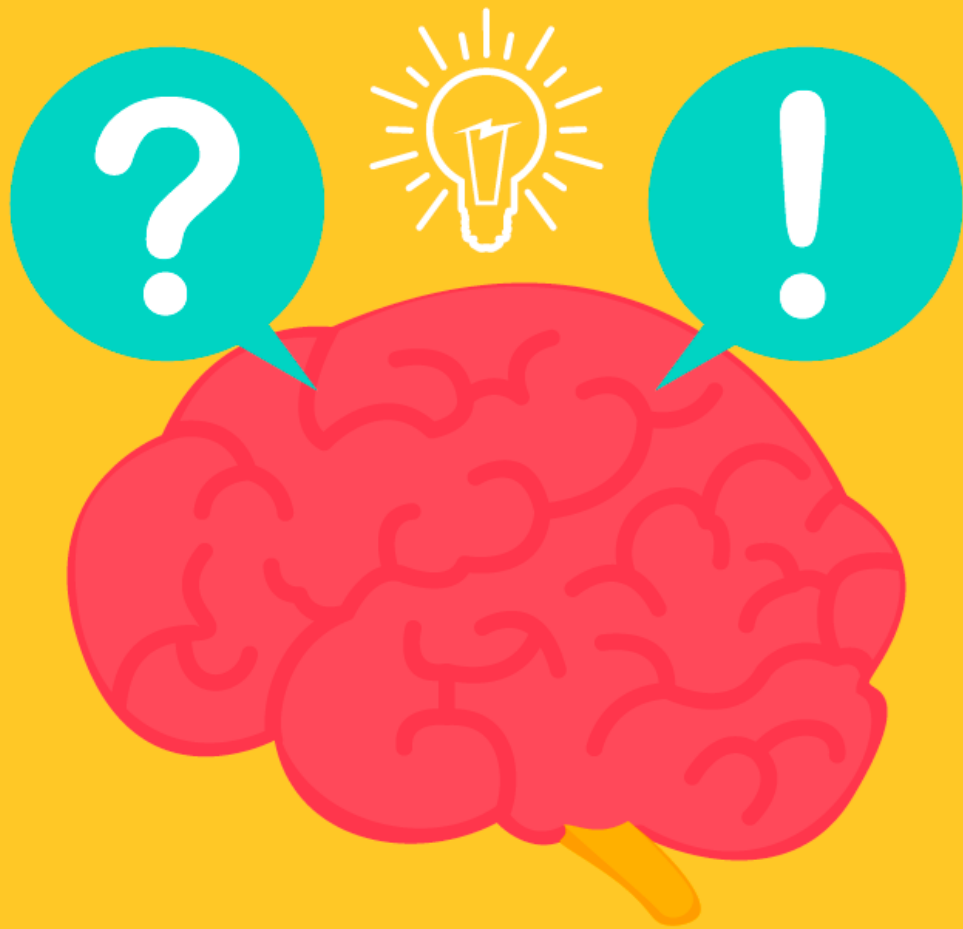


Figure 3. Mean truth ratings for true and false statements as a function of (A) the raw number of times viewed and (B) the natural logarithm of the number of views. Error bars reflect standard error of the mean. Participants responded on a scale of 1 = definitely false to 6 = definitely true.





Expect harmful beliefs caused by AI models to persist in the population even after you “fix the model”.

2.

2. AI models can't
“democratize knowledge”.

2. AI models can't
“democratize knowledge”.
They can't deliver equitable
results to all people because
people's language encodes
identity.



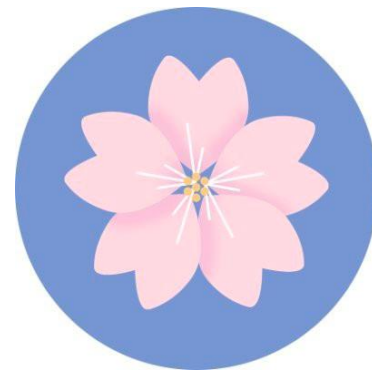
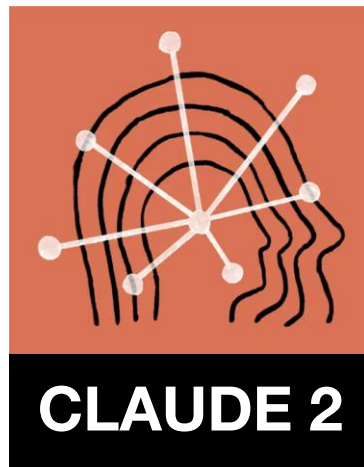


ANTHROPIC



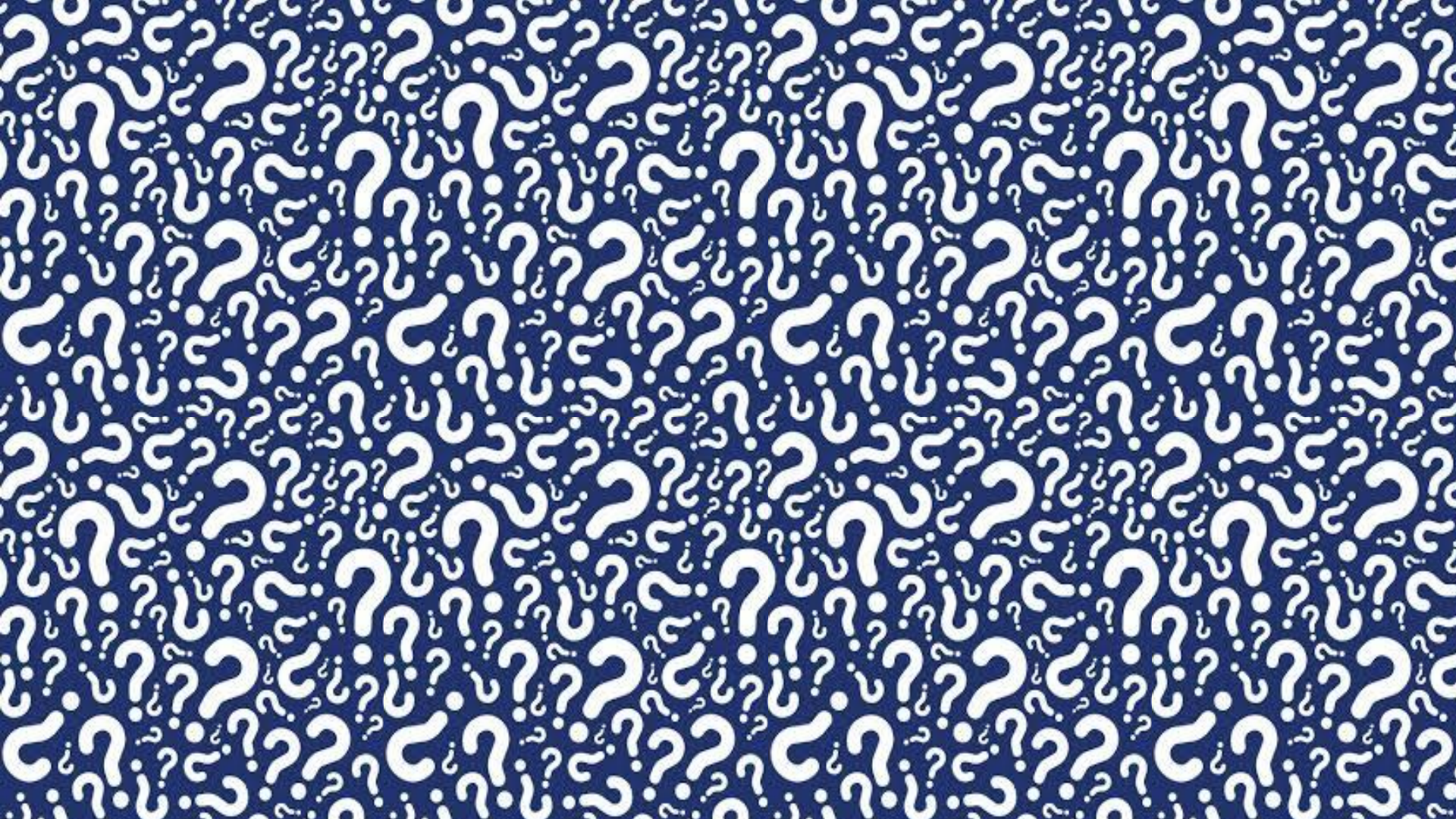
ChatGPT

**PaLM
2**



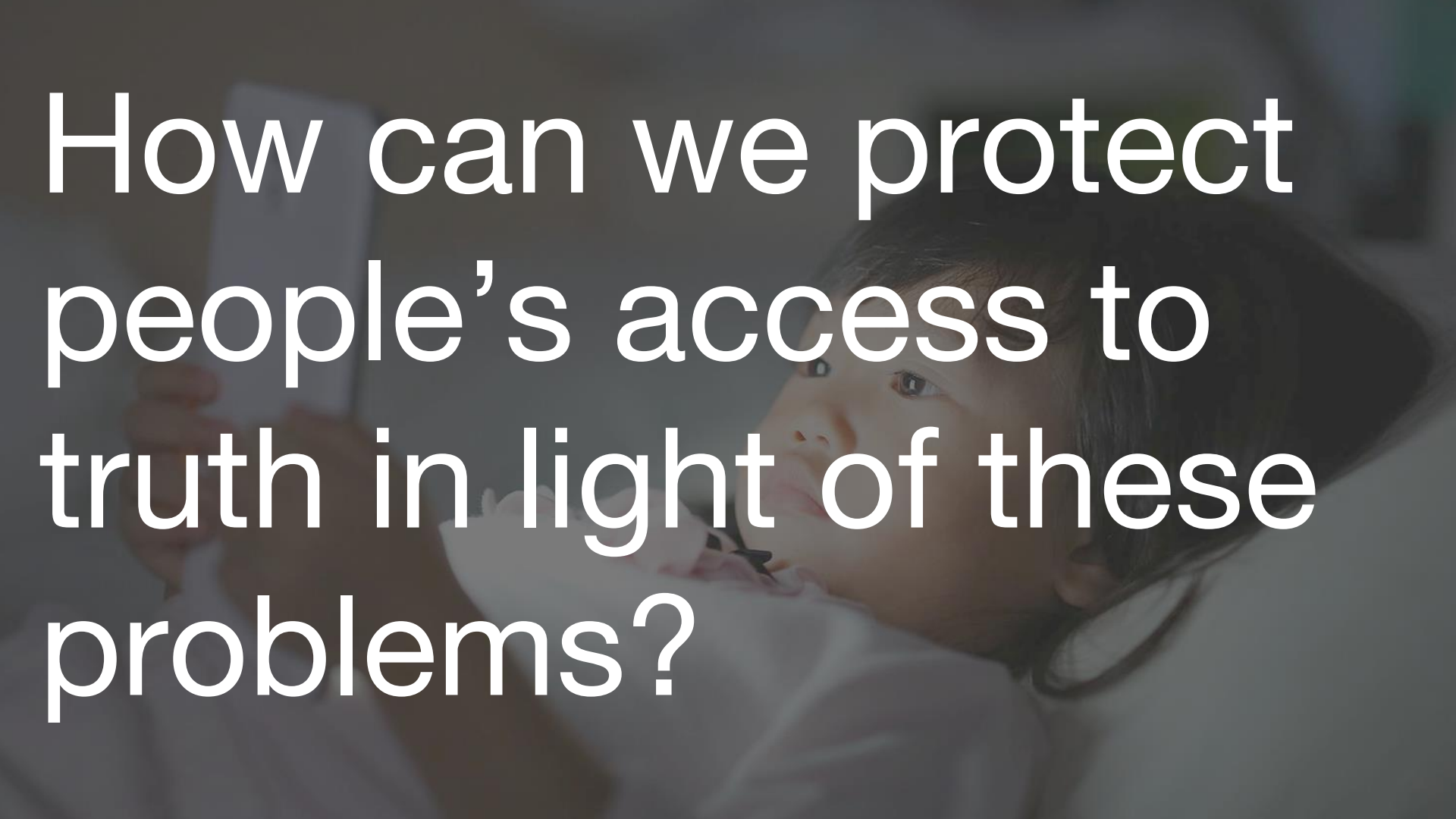
BLOOM





Two problematic ways AI distorts beliefs:

1. AI can transmit biases and fabrications to human users in ways that last.
2. AI models can't “democratize knowledge”.

A young child with dark hair is shown in profile, looking upwards and to the right with a thoughtful expression. The child is wearing a white shirt. The background is a soft, out-of-focus indoor setting. The text is overlaid on the image in a large, white, sans-serif font.

How can we protect
people's access to
truth in light of these
problems?

IEEE  CIS

CDS NEWSLETTER

The Newsletter of the Technical Committee on Cognitive and Developmental Systems

Volume 15, number 1
Spring 2018

Developmental Robotics
Machine Intelligence
Neuroscience
Psychology

Dialogue

Curiosity as Driver of Extreme Specialization in Humans



Celeste Kidd

Assistant Professor of
Psychology,
University of California,
Berkeley

celestekidd@gmail.com

The features that make us uniquely and distinctly human have been of interest to many people, from psychologists to philosophers to religious scholars, for centuries. Typical candidate traits include things like speech (Lieberman, 1991), upright posture (Clarke & Tobias, 1995), protracted childhoods (Jolly, 1972), helpless infants (Piantadosi & Kidd, 2016), sophisticated social cooperation (Melis & Semmann, 2010), and creativity (Carruthers, 2002).

There is, however, an essential human trait that has received far less recognition: the capacity for extreme specialization. Many humans spend a lifetime perfecting a single niche skill, such as a musical instrument, art medium, or style of dance. Others specialize in trades with economic roles (e.g., butchers, bakers, and candlestick makers). And while some other species exhibit certain forms of specialization—ants, for

entirely known or entirely novel ones (Dember & Earl, 1957; Kinney & Kagan, 1976; Berlyne, 1978; Kidd et al., 2013). More contemporary theories observe that curiosity is triggered when a gap is detected between what a learner currently knows, and what they could know (Loewenstein, 1994). This suggests the involvement of metacognition, since a learner must first identify that there is a gap to be filled before curiosity should be piqued. Yet little work to date has explored the relationship between metacognitive processes and curiosity. Are people who possess more metacognitive abilities pertaining to their own knowledge more curious? Can you make someone more curious by calling attention to what they do not know?

While we know that there exists some relationship between existing knowledge about a stimulus and the learner's degree of interest in

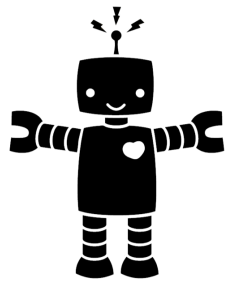


We have to preserve
diversity in human
experiences and
beliefs — and design AI
technologies with this
goal in mind.

Technologies that offer information to people should be sensitive to human psychology.



Berkeley
UNIVERSITY OF CALIFORNIA



Kidd Lab

How AI distorts human beliefs

Celeste Kidd (celestekidd@berkeley.edu)

Psychology, University of California, Berkeley



@celestekidd.bsky.social

World Summit AI, Doha, Qatar, 10 Dec 2024