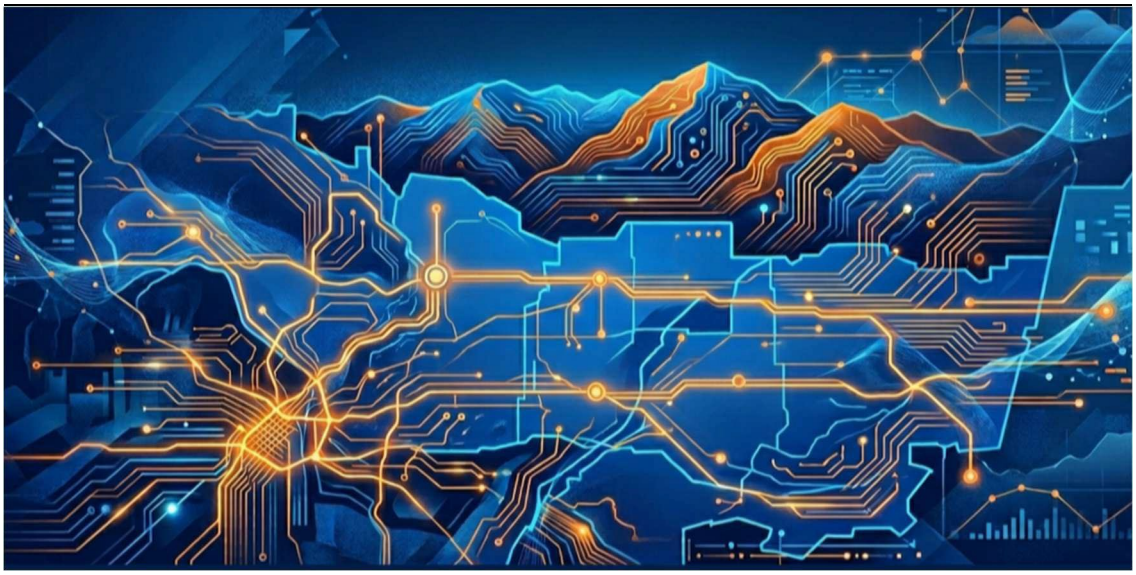
View this newsletter online.

LCSSP | The Ronald and Maxine Linde Center for Science, Society, and Policy

Quarterly Newsletter **July 2026**



Data Centers and Community Impact: A Workshop for Local Leaders

June 15, 2026

Hosted by the **Linde Center for Science, Society, and Policy (LCSSP)** in collaboration with the **Caltech Science Exchange** and the **Resnick Sustainability Institute**, this event brought community leaders together with researchers studying the local impacts of data centers, creating a space to share the latest findings, raising practical questions, and clarifying where important uncertainties remain. Just as importantly, this gathering—along with future convenings—will help ensure that research is informed by the real-world needs of communities in our region.

When communities can compare experiences, identify shared concerns, and exchange practical lessons, we are better positioned to make decisions that reflect shared values while benefitting from a wider base of knowledge.

Caltech's aim is not only to inform, but to build the relationships, questions, and shared understanding that can strengthen local capacity and open the door to more coordinated regional responses.

Learn more, including key insights and opportunities, in the **Caltech article** written on July 9, 2026. You can also watch a series of workshop videos on our **LCSSP YouTube channel**.

Announcements

James Mullahoo

Brenna Outten

Welcome to our first LCSSP Biopolicy Interns!

An important part of our mission is to empower Caltech students and postdocs with a better understanding of the interplay between scientific knowledge and government decision making with an opportunity for hands-on experience.

The Linde Science Policy Internship entails spending three months away from regular campus activities in order to complete a full-time internship working directly with a federal, state, or local government agency within the United States that is materially involved in science and technology issues.

By working alongside government officials, interns also have the opportunity to hone competencies related to effective communication, public engagement, interdisciplinary collaboration, and decision making under uncertainty.

Read more about our LCSSP interns and their host agencies:

- **James Mullahoo** - Maryland Department of Natural Resources, Chesapeake Bay Program, Submerged Aquatic Vegetation
- **Brenna Outten** - California Department of Developmental Services, Autism Services Branch

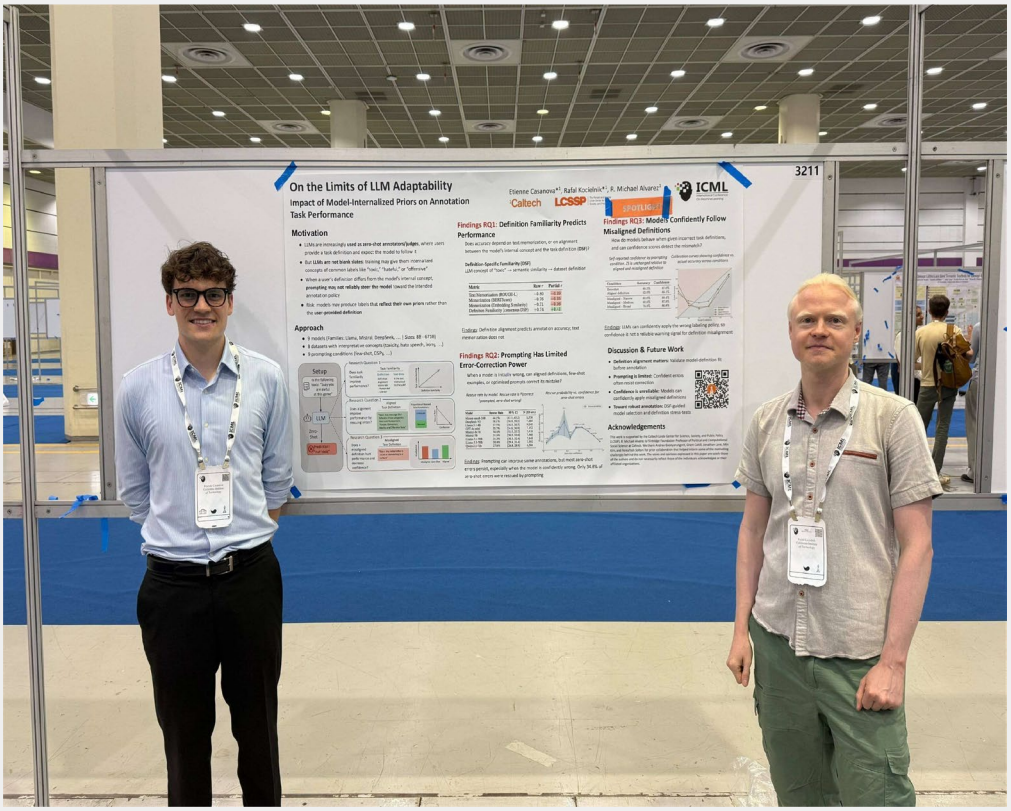
2026 SURF Students and WAVE Fellow

This summer we have three SURF students and one WAVE fellow at Caltech working with LCSSP researchers. Here is a brief introduction and a description of the projects they are working on:

- **Andrew Dou (Harvard)** - SURF student working with LCSSP Research Affiliates Jonathan Katz and Danny Ebanks on building an autoregressive Bayesian hierarchical model approximating the data generating process of U.S. House and Senate elections to quantify latent electoral quantities of interest. Andrew is also working on a second project with Jonathan Katz building a Bayesian hierarchical discrete two class mixture regression model to identify voters who have their vote for president significantly influenced by economic factors.
- **Zakiriya Gladney (Harvard)** - WAVE fellow working with LCSSP Postdoctoral Scholar Rafal Kocielnik and LCSSP Co-Director R. Michael Alvarez on a project that investigates whether manipulating the geometry of Large Language Models can help promote their behavioral consistency.
- **Shyna Kashi (Caltech)** - SURF student working with LCSSP Postdoctoral Scholar Raquel Centeno and LCSSP Co-Director R. Michael Alvarez to develop a Bayesian multilevel model that uses five decades of survey data to estimate how affective polarization has evolved across the United States.
- **Danielle Yang (Caltech)** - SURF student working with LCSSP Co-Director R. Michael Alvarez on analyzing the dimensionality of political podcast discourse and its potential collapse onto a left-right spectrum.

LCSSP Researchers Presenting at Academic Conferences

Several LCSSP researchers presented research at the International Conference on Machine Learning (ICML) on July 6-11, 2026 and the PolMeth Summer Meeting on July 15-18, 2026. Here is a listing of the authors and their research:



ICML 2026

- **Etienne Casanova**, **"On the Limits of LLM Adaptability: Impact of Model-Internalized Priors on Annotation Task Performance,"** with Rafal Kocielnik (Caltech) and R. Michael Alvarez (Caltech).
- **Rafal Kocielnik**, **"The Personality Illusion: Revealing Dissociation Between Self-Reports & Behavior in LLMs,"** with Pengrui Han (University of Illinois Urbana-Champaign), Peiyang Song (Carnegie Mellon University), Ramit Debnath (University of Cambridge), Dean Mobbs (Caltech), Anima Anandkumar (Caltech), and R. Michael Alvarez (Caltech).

PolMeth 2026

- **R. Michael Alvarez**, **"AI-Authored Prebunking Against Election Rumors: Evidence from Articles and Chatbots,"** with Mitchell Linegar (Caltech), Betsy Sinclair (WashU), and Sander van der Linden (University of Cambridge).
- **Raquel Centeno**, **"Who Trusted the 2024 Election? Assessing Voter Confidence in Ballot Counting Within and Across Racial and Ethnic Groups,"** with R. Michael Alvarez (Caltech).
- **Danny Ebanks**, "A Generatively Accurate, Unified Method of Evaluating Electoral Systems," with Jonathan N. Katz (Caltech) and Gary King (Harvard University).
- **Mitchell Linegar**, **"C-TreePO: Audited Compression Trees for Long-Document Measurement."**

Best Paper Award

Best Paper Award for joint paper at the ICML 2026 Conference in Seoul, Korea, July 6-11, 2026.

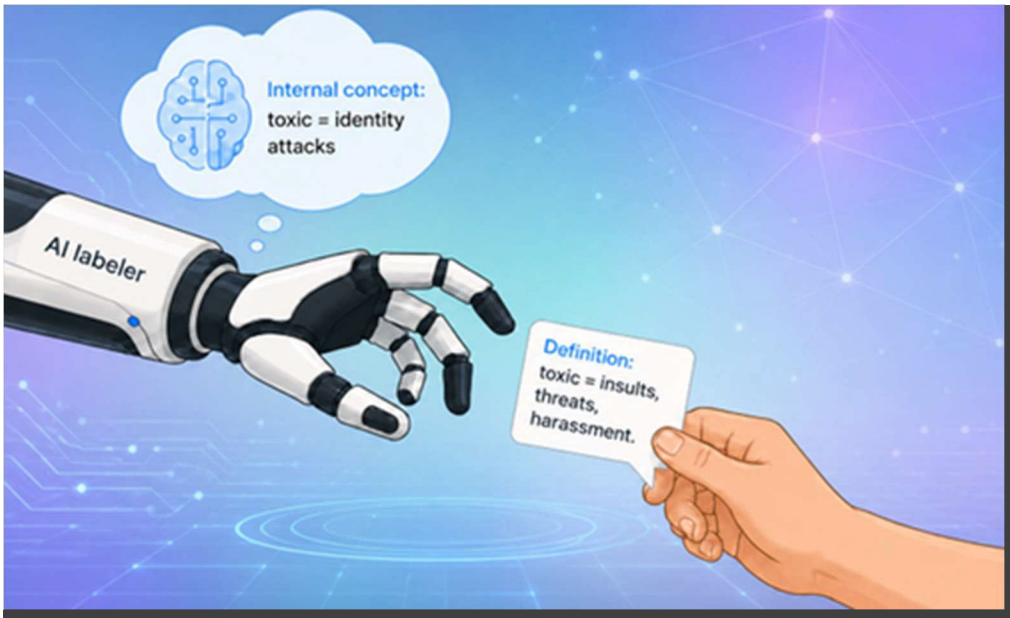
So exciting that joint paper, "Rethinking Psychometric Evaluation of LLMs: When and Why Self-Reports Predict Behavior," received the Best Paper Award at the **International Conference on Machine Learning** 2026 Workshop!

Big congratulations to lead author **Linde Center for Science, Society, and Policy (LCSSP) at Caltech** Postdoctoral Scholar **Rafal Kocielnik** (Caltech) and the entire team:

- **Pengrui Han**, University of Illinois Urbana-Champaign
- **Peiyang Song**, Carnegie Mellon University
- **Myrl G. Marmarelis**, Atmeto
- **Ramit Debnath**, University of Cambridge
- **Dean Mobbs**, Caltech
- **Anima Anandkumar**, Caltech
- **R. Michael Alvarez,** Linde Center for Science, Society, and Policy (LCSSP) at Caltech

To learn more about this collaboration, read the paper **here**.

New Research



The Hidden Assumptions Behind AI Labels

An investigation of how internalized beliefs shape AI annotation, and whether large language models follow human definitions or their own.

by Etienne Casanova, Rafal Kocielnik, and R. Michael Alvarez, California Institute of Technology

Large language models are quickly becoming "instant annotators." Instead of hiring teams of human data labelers, researchers can now ask an AI system to classify thousands of posts, comments, or survey responses in minutes. Platforms can use LLMs to flag harmful content. Social scientists can use them to code open-ended survey responses, interview transcripts, political speech, and other text that once required teams of trained coders. AI researchers can even use them as judges to evaluate other AI systems.

The appeal is obvious: give the model a definition, provide the text, and ask for a label.

But there is a hidden assumption behind this workflow: that the model will actually follow the definition we give it.

Our research asks a simple question with important consequences: when an LLM labels something as "toxic," "hateful," or "offensive," is it applying our definition, or one it has already learned?

Annotation depends on definitions

At first, a label like "toxic" might seem straightforward. But in practice, toxicity means different things in different settings.

In an online game, "you are awful at this" might be considered toxic because it insults another player and contributes to disruptive behavior. In a hate speech context, the same message might not be toxic because it does not target a protected identity group. In a news comment section, the boundary might be different again.

That means annotation is not just about matching text to labels. It is about applying a specific definition of a concept with contextual interpretation.

This is where LLMs become complicated. By the time you hand an LLM a rulebook, it has already absorbed enormous amounts of text during training. It has seen examples, discussions, moderation policies, and public debates about concepts like toxicity, hate speech, and offensiveness. Over time, it develops an internal understanding of these concepts.

The problem is that this internal understanding may not match the definition a researcher or platform wants to use. Worse, it may skew toward the majority interpretation, not the one a task requires.

Read more about their research on the **LCSSP Blog**.

Visit the LCSSP Website

Facebook X Instagram YouTube LinkedIn Bluesky

The LCSSP Newsletter is published by the **Linde Center for Science, Society, and Policy at Caltech.**

Our mailing address:

The Linde Center for Science, Society, and Policy

California Institute of Technology

Division of the Humanities and Social Sciences

1200 E. California Blvd.

Pasadena, CA 91125