

PANDORA 2026

**The First Workshop on Pluralistic AI & NLP:  
Diversity-aware, Sociotechnical, Responsible Alignment**

**Proceedings of the Workshop**

October 29, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)  
317 Sidney Baker St. S  
Suite 400 - 134  
Kerrville, TX 78028  
USA  
Tel: +1-855-225-1962  
[acl@aclweb.org](mailto:acl@aclweb.org)

ISBN 979-8-89176-484-2

## Preface

Welcome to the first PANDORA workshop, Pluralistic AI & NLP: Diversity-aware, Sociotechnical, Responsible Alignment, held on October 29, 2026, in conjunction with EMNLP 2026 in Budapest, Hungary. PANDORA brings together researchers and practitioners across natural language processing, machine learning, AI alignment, human–computer interaction, and the social sciences to examine how language technologies can account for diverse human values in a responsible way.

We invited empirical, methodological, conceptual, and interdisciplinary work on human label variation and pluralistic NLP; disagreement-aware annotation and modeling; pluralistic and preference-conditioned alignment; socialtechnical evaluation. Our aim is to connect technical approaches with careful accounts of their social setting and the people affected by them.

The first PANDORA Workshop received 92 submissions: 74 direct submissions and 18 ACL Rolling Review (ARR) commitments. In total, 30 papers were accepted, comprising 22 archival papers and 8 non-archival papers. This year’s workshop will also feature three invited talks.

We thank all authors for contributing their work and ideas, the reviewers for their careful and constructive assessments, the invited speakers for sharing their perspectives, and all participants for helping to build this community. We are also grateful to the EMNLP 2026 organizers for hosting the workshop and supporting its program.

The PANDORA 2026 Organizers

# Organizing Committee

## Organizers

Shanshan Xu, University of Copenhagen, Denmark  
Santosh T.Y.S.S., Amazon, USA  
Kemal Kurniawan, University of New South Wales, Australia  
Siyao Peng, LMU Munich, Germany  
Sarenne Wallbridge, AI Security Institute, UK  
Fajri Koto, MBZUAI, UAE  
Nouha Dziri, Allen Institute for AI (AI2), USA  
Barbara Plank, LMU Munich, Germany

## Program Committee

### Reviewers

Ahmed Abbasi, Zeeshan Ahmed

JinYeong Bak, Bal Krishna Bal, Valerio Basile, Nadjat Bouayad-Agha

Andrew Caines, Ilias Chalkidis, Beiduo Chen, Ming-Bin Chen, Juhwan Choi, Wenqian Cui

Ruchira Dhar, Yi Ding

Neele Falk, Diego Frassinelli

Yang Gao, Sanjana Gautam

Md. Mahadi Hassan, Yu Hou, Tianyi Hu, Zhengyu Hu

Pooja Jhunhunwala, Shan Jiang, Simeon Junker

Gyuwan Kim, Varun Kotte

Eric Le Ferrand, Tak Yeon Lee, Anqi Li, Alicia Liu

Maximilian Maurer, Selina Meyer, Kushan Mitra, Benedetta Muscato

Ninell Oldenburg

Jiaming Qu

Maria Ryskina

Danae Sanchez Villegas, Rita Sevastjanova, Sherrie Shen, Xiaoying Song, Leon Staufer, Kristina Striegnitz, Xiangguo Sun

Lei Wang, Yinggui Wang, Juan Wen

Kaige Xie

JungMin Yun

You Zhang, Zhang Zhang

# Keynote Talk

## TBD

**Vinodkumar Prabhakaran**  
Google Research  
**2026-10-29** – Room: **G13-14**

**Abstract:** To be announced.

**Bio:** Vinodkumar Prabhakaran is a Research Scientist at Google working at the intersection of natural language processing, machine learning, and social science. His research focuses on responsible AI, fairness, and cultural considerations in AI development and evaluation. He earned his PhD in Computer Science from Columbia University in 2015 and previously held a postdoctoral position at Stanford University.

# Keynote Talk

## A Psychological Perspective on Pluralistic AI Alignment

Xuechunzi Bai

University of Chicago

2026-10-29, 14:45–15:30 – Room: G13-14 / Online

**Abstract:** Alignment concerns how to ensure AI operates in accordance with human values. A key challenge is that human values vary across cultures and contexts, but this diversity is unevenly represented in the data from which models learn. Existing approaches primarily focus on putting guardrails to constrain model outputs without ensuring diversity in the data that shape model behavior. Drawing on psychology, I present three studies examining how biases persist, emerge, and can be mitigated. First, value-aligned models retain stereotypical associations that mirror human biases, showing that suppressing explicit bias does not necessarily eliminate implicit bias. Second, even when bias is absent initially, models can develop it by exploiting what has worked in their online experiences while insufficiently exploring alternatives. Their decisions shape subsequent experiences, reinforcing these biases and could compound them if incorporated into further training. Third, AI can help mitigate biases by nudging humans in producing more diverse content. Carefully designed AI writing assistants can encourage people to produce counter-stereotypical descriptions they might not otherwise consider. This suggests a way to diversify the representations that circulate in society and may become future training data. I argue that pluralistic alignment requires continual governance and human-centered intervention in the feedback loops linking models, people, and data, extending beyond one-off guardrails to the diversity of perspectives and experiences from which AI learns.

**Bio:** Xuechunzi Bai is the Neubauer Family Assistant Professor of Psychology at the University of Chicago. She is affiliated with the Center for Computational Social Science and the Center for Decision Research at the Booth School of Business. She studies how individual decisions and social structures interact to shape stereotypes and social biases in humans and AI systems. Her research combines large-scale behavioral experiments, computational cognitive modeling, naturalistic data analysis, and comparisons between humans and large language models. She received her PhD in Psychology and Social Policy from Princeton University, with a certificate in Statistics and Machine Learning in 2024.

# Keynote Talk

## To Personalize or not to Personalize? That is the Question

Vered Shwartz

University of British Columbia; Vector Institute  
2026-10-29, 16:00–16:45 – Room: G13-14 / Online

**Abstract:** Millions of users worldwide interact with LLMs daily to seek advice, information, and assistance. To maximize the effectiveness of these interactions, LLMs can maintain a memory of user preferences and context. In this talk, I argue that personalization carries both advantages and risks. I focus on cultural competence—the ability to respond in ways grounded in appropriate cultural knowledge and norms—as a lens for examining these trade-offs. I will first discuss (physical) localization, one form of personalization that can lead to more effective interactions, yet where current models still perform poorly. I will then argue for a more careful approach when generating conceptually different answers based on a user’s location or cultural background, as this risks perpetuating stereotypes and trapping users in echo chambers. Given that current models are heavily biased toward Western cultures, the status quo of one-size-fits-all is equally inadequate. To address this gap, I will present our work on LLMs with multiple cultural viewpoints.

**Bio:** Vered Shwartz is an Assistant Professor of Computer Science at the University of British Columbia, and a CIFAR AI Chair at the Vector Institute. Her research in natural language processing focuses on testing and improving the capabilities of large language models and vision and language models, developing culturally-competent AI, and responsible NLP applications in sensitive domains (e.g., legal, medical). Before joining UBC, she was a postdoctoral researcher at the Allen Institute for AI (AI2) and the Paul G. Allen School of Computer Science & Engineering at the University of Washington. Prior to that, she received her PhD in Computer Science from Bar-Ilan University in 2019.

## Table of Contents

|                                                                                                                                                                                                                                       |     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| <i>Demographic Pluralism: Inference-Time Modeling of Pluralistic Human Preference Distributions</i><br>Meng-Chen Wu, Qipin Chen, Ansh Jain, Tess Wood, Zhe Du and Si-Chi Chin .....                                                   | 1   |
| <i>The Grue Divide: When Cognitive Accuracy Doesn't Predict Affective Alignment</i><br>Kai Park .....                                                                                                                                 | 35  |
| <i>What is Cultural Representation" in AI Benchmarks? A Multidimensional Evaluation Framework</i><br>Carolyn Holtermann, Josefine Birkmann and Anne Lauscher .....                                                                    | 69  |
| <i>Metric-Dependent Annotation Saturation for Learning from Label Distributions</i><br>Guneet Kohli .....                                                                                                                             | 88  |
| <i>Pluralistic Annotation of Cultural Weaponization on Wikipedia: A Multi-Annotator Case Study across Ukraine and Armenia</i><br>Camil Hamdane, Hanna Perekhoda, Emanuela Boros, Stefanie Prezioso and Hamest Tamrazyan               | 105 |
| <i>Population-Grounded Evaluation of Multilingual LLM Alignment with European Societal Values</i><br>David Solans Noguero, Jordi Luque and Aleix Sant .....                                                                           | 121 |
| <i>When Context Sensitivity Isn't Enough: An Argument Against Likelihood-Based Probes for Evaluating LLM-Simulated Students</i><br>Pagnarith Pit and Ekaterina Vylomova .....                                                         | 135 |
| <i>Better Reward Models Are Not Always Less Biased: Social Alignment Beyond Standard Reward Benchmarks</i><br>Gayane Ghazaryan and Esra Dönmez .....                                                                                  | 149 |
| <i>Alignment Sharpens Elicited Label Distributions under Logit and Sampled Elicitation</i><br>Christy Jacob Tharakan and Kohsheen Tiku .....                                                                                          | 169 |
| <i>Which Objectives Need a Dial? Predicting Objective Conflict and Covering Trade-offs in Steerable Pluralistic Alignment</i><br>David Tsoi and Esra Dönmez .....                                                                     | 191 |
| <i>LatAm-RT: Toward Culturally Adaptive Red Teaming for AI Safety in Latin America</i><br>Roxana R. Romero and Minwoo Kim .....                                                                                                       | 207 |
| <i>PersonaWeaver: Controllable Diversity Beyond Conventional Archetypes in Procedural Character Generation</i><br>Maan Qraitem, Kate Saenko and Bryan A. Plummer .....                                                                | 224 |
| <i>Explain Less, Understand More: Data-Efficient Personalization of Reader-Dependent Jargons</i><br>Bohao Wu, Qingyun Wang and Yue Guo .....                                                                                          | 237 |
| <i>Evaluating with Disagreement: Metrics for Soft and Perspectivist Evaluation of Categorical and Ordinal Judgments</i><br>Giulia Rizzi, Silvia Casola, Elisabetta Fersini, Elisa Leonardelli, Barbara Plank and Massimo Poesio ..... | 250 |
| <i>Understanding Cultural Alignment in Multilingual LLMs via Natural Debate Statements</i><br>Vlad Andrei Negru, Camelia Lemnar, Mihai Surdeanu and Rodica Potolea .....                                                              | 266 |

|                                                                                                                                 |     |
|---------------------------------------------------------------------------------------------------------------------------------|-----|
| <i>Whose Toxicity? Persona Conditioning and Model–Annotator Label Agreement by Gender in Indonesian Toxicity Classification</i> |     |
| Salsabila Zahirah Pranida, Antonia Karamolegkou, Andrea Morales-Garzón and Daniel Hershcovich.....                              | 279 |
| <i>Persona Prompting in Multimodal Urban Perception: Descriptive Convergence and Interpretive Variation</i>                     |     |
| Neemias B da Silva, Matt Ratto, Myriam Delgado, Rodrigo Minetto, Daniel Silver and Thiago H. Silva.....                         | 301 |
| <i>Toward a Culture Map of Chatbots: Evaluating LLM Understanding of Cross-Cultural Workplace Norms</i>                         |     |
| Thomas Dubedout, Carlo Alberto Bono, Francesco Pierri and Anna Bernasconi.....                                                  | 319 |
| <i>Cultural Divergence Preservation: Diagnosing Flattening and Caricature in LLM-Simulated Survey Populations</i>               |     |
| Yeeun Chae, Yewon Choi, Seunghyun Lee and IL IM.....                                                                            | 334 |
| <i>BELIEFSIM: Towards Belief-Driven Simulation of Demographic Misinformation Susceptibility</i>                                 |     |
| Angana Borah, Zohaib Khan, Rada Mihalcea and Veronica Perez-Rosas.....                                                          | 348 |
| <i>Evaluating Cultural Awareness of LLMs for Haitian Creole</i>                                                                 |     |
| Christelle Clervilsson and Yanzhu Guo.....                                                                                      | 373 |
| <i>Plausible but Wrong: A Statistical Evaluation of GPT Models on High-Stakes Migration Statistics</i>                          |     |
| Hector Contreras, Ravinithesh Annapureddy and Daniel Gatica-perez.....                                                          | 388 |

# Program

## Thursday, October 29, 2026

|               |                                                    |
|---------------|----------------------------------------------------|
| 09:00 - 09:10 | <i>Opening</i>                                     |
| 09:10 - 09:50 | <i>Keynote Talk: <b>Vinodkumar Prabhakaran</b></i> |
| 09:50 - 10:40 | <i>Oral Session 1</i>                              |
| 10:40 - 11:10 | <i>Coffee Break</i>                                |
| 11:10 - 12:00 | <i>Oral Session 2</i>                              |
| 12:00 - 12:30 | <i>Poster Lightning Talk</i>                       |
| 12:30 - 13:30 | <i>Lunch</i>                                       |
| 13:30 - 14:45 | <i>Poster Session</i>                              |
| 14:45 - 15:30 | <i>Invited Talk: <b>Bai Xuechunzi</b></i>          |
| 15:30 - 16:00 | <i>Coffee Break</i>                                |
| 16:00 - 16:45 | <i>Invited Talk: <b>Vered Shwartz</b></i>          |
| 16:45 - 17:00 | <i>Best Paper and Closing</i>                      |