



THE PROGRAM GUIDE FOR THE  
**Seventh AAI/ACM Conference on**

# **ARTIFICIAL INTELLIGENCE, ETHICS, & SOCIETY**

**OCTOBER 21–23, 2024 | SAN JOSE, CA**

Follow AIES 2024 on X! @AIESConf



## AIES-24 WI-FI ACCESS

SSID

**SJCC Wi-Fi**

(No Password)

AIES-24 complimentary Wi-Fi is offered to support emails and web browsing. Please note, connections are limited to 4 mbps download and 2 mbps upload.

---



## TRAVELING TO SAN JOSE

 [Transportation to San Jose](#)

 [Parking Resources](#)

 [Interactive Parking Map](#)

 [Downtown Walking Map](#)

 [Nearby Dining Options](#)

## **PROGRAM CONTENTS**

---

<b>Acknowledgements</b>	<b>4</b>
<b>Welcome</b>	<b>5</b>
<b>Program Overview</b>	<b>7</b>
<b>Detailed Program</b>	<b>8</b>
Monday	8
Wednesday	14
Thursday	20
<b>Sponsors</b>	<b>24</b>



# ACKNOWLEDGEMENTS

The Association for the Advancement of Artificial Intelligence acknowledges and thanks the following individuals for their generous contributions of time and energy to the successful creation and planning of the Seventh Annual AAAI / ACM Conference on Artificial Intelligence, Ethics, and Society.

---

## CONFERENCE CO-CHAIRS

**Sanmay Das** *George Mason University*  
**Brian Patrick Green** *Santa Clara University*

---

## CONFERENCE PROGRAM CO-CHAIRS

**Kush Varshney** *IBM Research*  
**Marianna Ganapini** *Union College*  
**Andrea Renda** *Center for European Policy Studies*

---

## STUDENT PROGRAM CHAIRS

**Emanuelle Burton** *University of Illinois*  
**Yi Fang** *Santa Clara University*  
**Wenbin Zhang** *Florida International University, USA*

# The AIES-24 Program Committee welcomes you to San Jose!



## WELCOME FROM THE CONFERENCE CO-CHAIRS

The mission of AIES is to engage a multidisciplinary group of scholars to think deeply about the impact of AI systems on human societies. We are thrilled that the conference is growing while still maintaining very high quality standards and the transdisciplinary nature that is so core to its identity. We are very excited at the tremendous program that AIES has to offer its attendees this year (and, in perpetuity, those who will read the papers in the proceedings). We look forward to welcoming all of you to San Jose and hope you enjoy seeing the incredible work that is taking place in the field. We are grateful to our sponsors for their generous support, which enables us to keep registration fees low and support the student program. We are at a critical moment in time, as AI becomes increasingly pervasive. It is our hope that the conversations at AIES continue to drive the work we need to do to ensure that the path forward is a good one.

*Sanmay Das* *George Mason University* · *Brian Patrick Green* *Santa Clara University*



## WELCOME FROM THE CONFERENCE PROGRAM CO-CHAIRS

Welcome to the 7th annual AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society in the heart of Silicon Valley! We are grateful to the support from our hosts at the Markkula Center for Applied Ethics at Santa Clara University, and the innumerable hours put in by Chesley Grove, Meredith Ellison, the rest of the AAAI staff, and the general chairs in arranging the event so that we could focus on bringing you a high-quality technical program.

AIES has grown from a small gathering in 2018 co-located with the main AAAI conference with 57 papers to a selective, full-fledged, independent conference where a trans-disciplinary collection of researchers gathered to learn from each other. With rapidly increasing capabilities of AI, AI's permeation in increasing realms of life, and increasing regulations and voluntary commitments on AI safety, there is no more important time for us to convene than now.

This year, after desk-rejecting a handful of submissions, we sent 468 papers out for review and accepted 150 of them; hands down, both are record numbers. We are

extremely thankful to the 219 program committee members from all over the world who volunteered their time to provide peer reviews. The papers will be presented through a combination of oral sessions and poster sessions. One form of presentation is not more prestigious than the other; the session assignments were made to have a coherent and engaging program within the constraints of two-and-a-half days. We are also pleased to have three distinguished keynote speakers: danah boyd, David Danks, and Diyi Yang.

We hope you enjoy the conference and walk away with new friends and new ideas.

**Kush Varshney** *IBM Research* · **Marianna Ganapini** *Union College*  
**Andrea Renda** *Center for European Policy Studies*



## **WELCOME FROM THE STUDENT PROGRAM CHAIRS**

Every year, the AIES Student Program offers support to exceptional students from all over the world, helping them make the trip to AIES and connecting them to the vibrant interdisciplinary community of scholars that gathers here. These students will be presenting their work at the poster sessions each day. We hope you'll make time to talk to several of them and hear about the wide variety of exciting projects they are pursuing.

We are grateful to NSF, SIGAI, and AAI for financial support, and also grateful to those senior scholars who are serving as mentors to the students.

**Emanuelle Burton** *University of Illinois* · **Yi Fang** *Santa Clara University*  
**Wenbin Zhang** *Florida International University, USA*



# PROGRAM OVERVIEW

## MONDAY, OCTOBER 21

9:00  
Opening Remarks &  
Paper Awards

---

9:15  
Oral Session 1

---

10:15  
Break

---

10:45  
Keynote 1

---

11:45  
Oral Session 2

---

12:45  
Lunch Break

---

2:15  
Oral Session 3

---

3:15  
Break

---

3:45  
Oral Session 4

---

4:45  
Break

---

5:00  
Oral Session 5

---

6:00  
Poster Session 1 & Reception

## TUESDAY, OCTOBER 22

9:00  
Pedagogy Roundtable

---

10:15  
Break

---

10:45  
Keynote 2

---

11:45  
Oral Session 6

---

12:45  
Lunch Break

---

2:15  
Oral Session 7

---

3:15  
Break

---

3:45  
Oral Session 8

---

4:45  
Break

---

5:00  
Oral Session 9

---

6:00  
Poster Session 2 & Reception

## WEDNESDAY, OCTOBER 23

9:00  
Poster Session 3 &  
Breakfast

---

10:45  
Oral Session 10

---

11:45  
Break

---

12:15  
Keynote 3

---

1:15  
Closing Remarks

# MONDAY, OCTOBER 21

- 9:00–9:10**    **Opening Remarks & Paper Awards // Room LL20AB**
- 9:15–10:15**    **Oral Session 1 – Algorithmic Implications of Regulations // Room LL20AB**
- 120**            How Should AI Decisions Be Explained?  
Requirements for Explanations from the Perspective of European Law  
*Benjamin Fresz, Elena Dubovitskaya, Danilo Brajovic, Marco Huber and Christian Horz*
- 215**            Proxy Fairness under the European Data Protection Regulation and the AI ACT:  
A Perspective of Sensitivity and Necessity  
*Ioanna Papageorgiou*
- 203**            You Still See Me: How Data Protection Supports the Architecture of AI Surveillance  
*Rui-Jie Yew, Lucy Qin and Suresh Venkatasubramanian*
- 341**            Do Responsible AI Artifacts Advance Stakeholder Goals?  
Four Key Barriers Perceived by Legal and Civil Stakeholders  
*Anna Kawakami, Daricia Wilkinson and Alexandra Chouldechova*
- 10:15–10:45**    **Break**
- 10:45–11:45**    **Keynote 1 // Room LL20AB**  
Diyi Yang, Stanford University
- 11:45–12:45**    **Oral Session 2 – Large Language Model Alignment // Room LL20AB**
- 111**            Learning When Not to Measure: Theorizing Ethical Alignment in LLMs  
*William Rathje*
- 256**            A Qualitative Study on Cultural Hegemony and the Impacts of AI  
*Venetia Brown, Retno Larasati, Aisling Third and Tracie Farrell*
- 455**            PoliTune: Analyzing the Impact of Data Selection and Fine-Tuning on Economic  
and Political Biases in Large Language Models  
*Ahmed Agiza, Mohamed Mostagir and Sherief Reda*
- 453**            Legal Minds, Algorithmic Decisions: How LLMs Apply Constitutional Principles in Complex Scenarios  
*Carolina Camassa and Camilla Bignotti*



**MONDAY, OCTOBER 21****12:45–2:15**    **Lunch Break (on own)****2:15–3:15**    **Oral Session 3 – Excluded Knowledges and Openness // Room LL20AB**

- 150**    What Makes An Expert? Reviewing How ML Researchers Define “Expert”  
*Mark Diaz and Angela Smith*
- 57**    Surveys Considered Harmful? Reflecting on the Use of Surveys in AI Research, Development, and Governance  
*Mohammad Tahaei, Daricia Wilkinson, Alisa Frik, Michael Muller, Ruba Abu-Salma and Lauren Wilcox*
- 39**    Decolonial AI Alignment: Openness, Visesa–Dharma, and Including Excluded Knowledges  
*Kush R. Varshney*
- 66**    The Origin and Opportunities of Developers’ Perceived Code Accountability in Open Source AI Software Development  
*Sebastian Clemens Bartsch, Moritz Lothar, Jan-Hendrik Schmidt, Martin Adam and Alexander Benlian*

**3:15–3:45**    **Break****3:45–4:45**    **Oral Session 4 – Governance and Implications // Room LL20AB**

- 40**    Pay Attention: a Call to Regulate the Attention Market and Prevent Algorithmic Emotional Governance  
*Franck Michel and Fabien Gandon*
- 492**    Acceptable Use Policies for Foundation Models  
*Kevin Klyman*
- 227**    An FDA for AI? Pitfalls and Plausibility of Approval Regulation for Frontier Artificial Intelligence  
*Daniel Carpenter and Carson Ezell*
- 336**    The Societal Implications of Open Generative Models Through the Lens of Fact-Checking Organizations  
*Robert Wolfe and Tanushree Mitra*

**4:45–5:00**    **Break****5:00–6:00**    **Oral Session 5 – Responsible AI Tools and Transparency // Room LL20AB**

- 228**    Foundation Model Transparency Reports  
*Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan and Percy Liang*
- 447**    Co-designing an AI Impact Assessment Report Template with AI Practitioners and AI Compliance Experts  
*Edyta Bogucka, Marios Constantinides, Sanja Scepanovic and Daniele Quercia*

**MONDAY, OCTOBER 21**

23 How Do AI Companies “Fine-Tune” Policy? Examining Regulatory Capture in AI Governance  
*Kevin Wei, Carson Ezell, Nick Gabrieli and Chinmay Deshpande*

397 The Ethico-Politics of Design Toolkits:  
Responsible AI Tools, From Big Tech Guidelines to Feminist Ideation Cards  
*Tomasz Hollanek*

**6:00–8:00 Poster Session 1 & Reception // Room LL20CD**

**Posters**

28 What’s Distributive Justice Got to Do with It? Rethinking Algorithmic Fairness from a  
Perspective of Approximate Justice  
*Corinna Hertweck, Christoph Heitz and Michele Loi*

43 Reflection of its Creators: Qualitative Analysis of General Public  
and Expert Perceptions of Artificial Intelligence  
*Theodore Jensen, Mary Theofanos, Kristen Greene, Olivia Williams, Kurtis Goad and Janet Bih Fofang*

48 Habemus a Right to an Explanation: so What? –  
A Framework on Transparency–Explainability Functionality and Tensions in the EU AI Act  
*Luca Nannini*

77 A Formal Account of Trustworthiness: Connecting Intrinsic and Perceived Trustworthiness  
*Piercosma Bisconti, Letizia Aquilino, Antonella Marchetti and Daniele Nardi*

80 Quantifying gendered citation imbalance in computer science conferences  
*Kazuki Nakajima, Yuya Sasaki, Sohei Tokuno and George Fletcher*

93 PPS: Personalized Policy Summarization for Explaining Sequential Behavior of Autonomous Agents  
*Peizhu Qian, Harrison Huang and Vaibhav V. Unhelkar*

101 On the Trade-offs between Adversarial Robustness and Actionable Explanations  
*Satyapriya Krishna, Chirag Agarwal and Himabindu Lakkaraju*

108 “I don’t see myself represented here at all”: User Experiences of Stable Diffusion Outputs  
Containing Representational Harms across Gender Identities and Nationalities  
*Sourojit Ghosh, Nina Lutz and Aylin Caliskan*

123 Racial and Neighborhood Disparities in Legal Financial Obligations in Jefferson County, Alabama  
*Óscar Lara Yejas, Aakanksha Joshi, Andrew Martinez, Leah Nelson, Skyler Speakman,  
Krysten Thompson, Yuki Nishimura, Jordan Bond and Kush R. Varshney*

134 Estimating Environmental Cost Throughout Model’s Adaptive Life Cycle  
*Vishwesh Sangarya, Richard Bradford and Jung-Eun Kim*

136 The Impact of Responsible AI Research on Innovation and Development  
*Ali Septiandri, Marios Constantinides and Daniele Quercia*

# MONDAY, OCTOBER 21

- 143 Public Attitudes on Performance for Algorithmic and Human Decision-Makers  
*Kirk Bansak and Elisabeth Paulson*
- 174 “Democratizing AI” and the Concern of Algorithmic Injustice  
*Ting-An Lin*
- 178 Kid-Whisper: Towards Bridging the Performance Gap in Automatic Speech Recognition for Children VS. Adults  
*Ahmed Attia, Jing Liu, Wei Ai, Dora Demszky and Carol Espy-Wilson*
- 213 CIVICS: Building a Dataset for Examining Culturally-Informed Values in Large Language Models  
*Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni and Margaret Mitchell*
- 233 Introducing the AI Governance and Regulatory Archive (AGORA): An Analytic Infrastructure for Navigating the Emerging AI Governance Landscape  
*Zachary Arnold, Niharika Singh, Jennifer Melot, Daniel S. Schiff, Kaylyn Jackson Schiff, Ogadinma Enweareazu and Tyler Girard*
- 241 Enhancing Equitable Access to AI in Housing and Homelessness System of Care through Federated Learning  
*Musa Taib, Jiajun Wu, Steve Drew and Geoffrey Messier*
- 258 The Problems with Proxies: Making Data Work Visible through Requester Practices  
*Annabel Rothschild, Ding Wang, Niveditha Jayakumar, Lauren Wilcox, Carl DiSalvo and Betsy DiSalvo*
- 261 Scaling Laws Do Not Scale  
*Fernando Diaz and Michael Madaio*
- 285 Beyond Thumbs Up/Down: Untangling Challenges of Fine-Grained Feedback for Text-to-Image Generation  
*Katherine Collins, Najoung Kim, Yonatan Bitton, Verena Rieser, Shayegan Omidshafiei, Yushi Hu, Sherol Chen, Senjuti Dutta, Minsuk Chang, Kimin Lee, Youwei Liang, Georgie Evans, Sahil Singla, Gang Li, Adrian Weller, Junfeng He, Deepak Ramachandran and Krishnamurthy Dvijotham*
- 287 Compassionate AI for Moral Decision-Making, Health, and Well-Being  
*Mark Graves and Jane Compson*
- 288 Formal Ethical Obligations in Reinforcement Learning Agents: Verification and Policy Updates  
*Colin Shea-Blymyer and Houssam Abbas*
- 296 Simulating Policy Impacts: Developing a Generative Scenario Writing Method to Evaluate the Perceived Effects of Regulation  
*Julia Barnett, Kimon Kieslich and Nicholas Diakopoulos*
- 306 Contributory injustice, epistemic calcification and the use of AI systems in healthcare  
*Mahi Hardalupas*
- 334 Mitigating urban-rural disparities in contrastive representation learning with satellite imagery  
*Miao Zhang and Rumi Chunara*
- 339 ML-EAT: A Multilevel Embedding Association Test for Interpretable and Transparent Social Science  
*Robert Wolfe, Alexis Hiniker and Bill Howe*

- 345 Observing Context Improves Disparity Estimation when Race is Unobserved  
*Kweku Kwegyir-Aggrey, Naveen Durvasula, Jennifer Wang and Suresh Venkatasubramanian*
- 371 Algorithmic Fairness From the Perspective of Legal Anti-discrimination Principles  
*Vijay Keswani and L. Elisa Celis*
- 389 Reducing Biases towards Minoritized Populations in Medical Curricular Content via Artificial Intelligence for Fairer Health Outcomes  
*Chiman Salavati, Shannon Song, Willmar Sosa Diaz, Scott A. Hale, Roberto E. Montenegro, Fabricio Murai and Shiri Dori-Hacohen*
- 395 Tracing the Evolution of Information Transparency for OpenAI's GPT Models Through a Biographical Approach  
*Zhihan Xu and Eniana Mustafaraj*
- 415 MoJE: Mixture of Jailbreak Experts, Naive Tabular Classifiers as Guard for Prompt Attacks  
*Giandomenico Cornacchia, Kieran Fraser, Muhammad Zaid Hameed, Mark Purcell, Amrisha Rawat and Giulio Zizzo*
- 417 Social Scoring Systems for Behavioral Regulation: An Experiment on the Role of Transparency in Determining Perceptions and Behaviors  
*Carmen Löfflad, Mo Chen and Jens Grossklags*
- 420 Trustworthy Social Bias Measurement  
*Rishi Bommasani and Percy Liang*
- 434 A Conceptual Framework for Ethical Evaluation of Machine Learning Systems  
*Neha Gupta, Jessica Hullman and Hariharan Subramonyam*
- 442 The PPOu Framework: A Structured Approach for Assessing the Likelihood of Malicious Use of Advanced AI Systems  
*Josh A. Goldstein and Girish Sastry*
- 450 Epistemic Injustice in Generative AI  
*Jackie Kay, Atoosa Kasirzadeh and Shakir Mohamed*
- 451 Disengagement through Algorithms: How Traditional Organizations Aim for Experts' Satisfaction  
*Jérémie Poiroux*
- 461 Decoding Multilingual Moral Preferences: Unveiling LLM's Biases Through the Moral Machine Experiment  
*Karina Vida, Fabian Damken and Anne Lauscher*
- 465 Stable Diffusion Exposed: Gender Bias from Prompt to Image  
*Yankun Wu, Yuta Nakashima and Noa Garcia*
- 472 Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents  
*Elizaveta Tennant, Stephen Hailes and Mirco Musolesi*
- 414 LLM Platform Security: Applying a Systematic Evaluation Framework to OpenAI's ChatGPT Plugins  
*Umar Iqbal, Tadayoshi Kohno and Franziska Roesner*

## *Student Posters*

- 29      Untangling Race and Lung Function in Prediction Models  
*Amin Adibi*
  
- 24      Fair Conversational Recommender System  
*Mina Arzaghi*
  
- 33      Misplaced Capabilities: Evaluating the Risks of Anthropomorphism in Human-AI Interactions  
*Takuya Maeda*
  
- 21      Tree-Based Approaches for Interpretable Modeling in Healthcare  
*Juliette Murriss*
  
- 17      Advancing Early Alzheimer's Disease Detection in Underdeveloped Areas  
with Fair Explainable AI Methods  
*Quoc-Toan Nguyen*
  
- 34      An Intersectional Approach to Large Language Models  
*Khaoula Chehbouni*
  
- 20      Leveraging Mixed Methods to Identify and Address Technological Harms towards Marginalized Groups  
*Camille Harris*
  
- 15      Uncovering Gender Biases in Human-AI Platforms  
*Siddarth Jaiswal*
  
- 1      Algorithmic Decision-Making under Agents with Persistent Improvement  
*Tian Xie*
  
- 28      Generative Models for Art and Society  
*Yankun Wu*

# TUESDAY, OCTOBER 22

- 9:00–10:15**    **Pedagogy Roundtable // Room LL20AB**  
Emanuelle Burton, University of Illinois Chicago  
Casey Fiesler, University of Colorado Boulder  
Amy J. Ko, University of Washington Information School  
Amanda McCroskery, Google Deepmind  
Marty J. Wolf, Bemidji State University
- 10:15–10:45**    **Break**
- 10:45–11:45**    **Keynote 2 // Room LL20AB**  
David Danks, University of California San Diego
- 11:45–12:45**    **Oral Session 6 – Biases in Foundation Models I // Room LL20AB**
- 317**            Examining the Behavior of LLM Architectures Within the Framework of Standardized National Exams in Brazil  
*Marcelo Sartori Locatelli, Matheus Prado Miranda, Igor Joaquim Costa, Matheus Torres Prates, Victor Thome, Mateus Zaparoli, Tomas Lacerda, Adriana Pagano, Eduardo Rios Neto, Wagner Meira Jr. and Virgilio Almeida*
- 359**            Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval  
*Kyra Wilson and Aylin Caliskan*
- 8**                A Causal Framework to Evaluate Racial Bias in Law Enforcement Systems  
*Jessy Xinyi Han, Andrew Miller, S. Craig Watkins, Christopher Winship, Fotini Christia and Devavrat Shah*
- 323**            Automate or Assist? The Role of Computational Models in Identifying Gendered Discourse in US Capital Trial Transcripts  
*Andrea W Wen-Yi, Kathryn Adamson, Nathalie Greenfield, Rachel Goldberg, Sandra Babcock, David Mimno and Allison Koenecke*
- 12:45–2:15**    **Lunch Break (on own)**
- 2:15–3:15**    **Oral Session 7 – Human-AI Relationships // Room LL20AB**
- 79**              Perception of experience influences altruism and perception of agency influences trust in human-machine interactions  
*Mayada Oudah, Kinga Makovi, Kurt Gray, Balaraju Battu and Talal Rahwan*
- 367**              What Is Required for Empathic AI? It Depends, and Why That Matters for AI Developers and Users  
*Jana Schaich Borg and Hannah Read*

**TUESDAY, OCTOBER 22**

- 200 Beyond Interaction: Investigating the Appropriateness of Human–AI Assistant Relationships  
*Arianna Manzini, Geoff Keeling, Lize Alberts, Shannon Vallor, Meredith Ringel Morris and Iason Gabriel*
- 234 Unsocial Intelligence: an Investigation of the Assumptions of AGI Discourse  
*Borhane Blili-Hamelin, Leif Hancox-Li and Andrew Smart*
- 3:15–3:45 Break**
- 3:45–4:45 Oral Session 8 – Algorithms // Room LL20AB**
- 86 Fairness in Reinforcement Learning: A Survey  
*Anka Reuel and Devin Ma*
- 496 Nothing Comes Without Its World – Practical Challenges of Aligning LLMs to Situated Human Values through RLHF  
*Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon and Jie Yang*
- 32 Algorithmic Decision-Making under Agents with Persistent Improvement  
*Tian Xie, Xuwei Tan and Xueru Zhang*
- 487 When and Why is Persuasion Hard? A Computational Complexity Result  
*Zach Wojtowicz*
- 4:45–5:00 Break**
- 5:00–6:00 Oral Session 9 – Evaluating Risks and Harms // Room LL20AB**
- 418 ExploreGen: Large Language Models for Envisioning the Uses and Risks of AI Technologies  
*Viviane Herdel, Sanja Šćepanovic, Edyta Bogucka and Daniele Quercia*
- 147 Red-Teaming for Generative AI: Silver Bullet or Security Theater?  
*Michael Feffer, Anusha Sinha, Wesley Deng, Zachary Lipton and Hoda Heidari*
- 146 Gaps in the Safety Evaluation of Generative AI  
*Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac and Laura Weidinger*
- 433 Operationalizing content moderation “accuracy” in the Digital Services Act  
*Johnny Wei, Frederike Zufall and Robin Jia*

**TUESDAY, OCTOBER 22****6:00–8:00 Poster Session 2 & Reception // Room LL20CD****Posters**

- 26 Strategies for Increasing Corporate Responsible AI Prioritization  
*Angelina Wang, Teresa Datta and John Dickerson*
- 29 On Feasibility of Intent Obfuscating Attacks  
*Zhaobin Li and Patrick Shafto*
- 31 Non-linear Welfare-Aware Strategic Learning  
*Tian Xie and Xueru Zhang*
- 45 Gender in pixels: pathways to non-binary representation in Computer Vision  
*Elena Beretta*
- 60 Responsible Reporting for Frontier AI Development  
*Noam Kolt, Markus Anderljung, Joslyn Barnhart, Asher Brass, Kevin Esvelt, Gillian K. Hadfield, Lennart Heim, Mikel Rodriguez, Jonas B. Sandbrink and Thomas Woodside*
- 64 Coordinated Disclosure for AI: Beyond Security Vulnerabilities  
*Sven Cattell, Avijit Ghosh and Lucie-Aimée Kaffee*
- 89 The supply chain capitalism of AI: A call to (re)think algorithmic harms and resistance  
*Ana Valdivia*
- 95 APPRAISE: a Governance Framework for Innovation with Artificial Intelligence Systems  
*Diptish Dey and Debarati Bhaumik*
- 106 Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation  
*Declan Grabb, Max Lamparth and Nina Vasan*
- 110 Individual Fairness in Graphs Using Local and Global Structural Information  
*Yonas Sium, Qi Li and Kush R. Varshney*
- 117 A Human-in-the-Loop Fairness-Aware Model Selection Framework for Complex Fairness Objective Landscapes  
*Jake Robertson, Thorsten Schmidt, Frank Hutter and Noor Awad*
- 118 Algorithms and Recidivism: A Multi-disciplinary Systematic Review  
*Arul Scaria, Vidya Subramanian, Nevin George and Nandana Sengupta*
- 124 Introducing ELLIPS: an ethics-centered approach to research on LLM-based inference of psychiatric conditions  
*Roberta Rocca, Giada Pistilli, Kritika Maheshwari and Riccardo Fusaroli*
- 137 Face the Facts: Using Face Averaging to Visualize Gender-by-Race Bias in Facial Analysis Algorithms  
*Kentrell Owens, Erin Freiburger, Ryan Hutchings, Mattea Sim, Kurt Hugenberg, Franziska Roesner and Tadayoshi Kohno*



**TUESDAY, OCTOBER 22**

- 145 A Relational Justification of AI Democratization  
*Bauke Wielinga and Stefan Buijsman*
- 153 SoUnD Framework: Analyzing (So)cial Representation in (Un)structured (D)ata  
*Mark Diaz, Sunipa Dev, Emily Reif, Emily Denton and Vinodkumar Prabhakaran*
- 163 What's Your Stake in Sustainability of AI?: An Informed Insider's Guide  
*Grace C. Kim, Annabel Rothschild, Carl DiSalvo and Betsy DiSalvo*
- 165 AI debates aren't binary – they're plural  
*Thorin Bristow, Diana Acosta Navas and Luke Thorburn*
- 175 Do Generative AI Models Output Harm while Representing Non-Western Cultures: Evidence from A Community-Centered Approach  
*Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson and Aylin Caliskan*
- 189 Beyond Participatory AI  
*Jonne Maas and Aarón Moreno Inglés*
- 196 PICE: Polyhedral Complex Informed Counterfactual Explanations  
*Mattia Jacopo Villani, Emanuele Albini, Shubham Sharma, Saumitra Mishra, Salim Ibrahim Amoukou, Daniele Magazzeni and Manuela Veloso*
- 201 Fairness in AI-Based Mental Health: Clinician Perspectives and Bias Mitigation  
*Gizem Sogancioglu, Pablo Mosteiro, Albert Ali Salah, Floortje Scheepers and Heysem Kaya*
- 244 Foundations for Unfairness in Anomaly Detection – Case Studies in Facial Imaging Data  
*Michael Livanos and Ian Davidson*
- 247 Medical AI, Categories of Value Conflict, and Conflict Bypasses  
*Gavin Victor and Jean-Christophe Bélisle-Pipon*
- 251 Lessons from clinical communications for AI systems  
*Alka Menon, Zahra Abba Omar, Nadia Nahar, Xenophon Papademetris, Lynn Fiellin and Christian Kästner*
- 253 Virtual Assistants Are Unlikely to Reduce Patient Non-Disclosure  
*Corinne Jorgenson, Ali Ozkes, Jurgen Willems and Dieter Vanderelst*
- 267 Uncovering the gap: Challenging the Agential Nature of AI Responsibility Problems  
*Joan Llorca Albareda*
- 275 Hidden or Inferred: Fair Learning-To-Rank With Unknown Demographics  
*Oluseun Olulana, Kathleen Cachel, Fabricio Murai and Elke Rundensteiner*
- 280 LLMs and Memorization: On Quality and Specificity of Copyright Compliance  
*Felix Benjamin Müller, Rebekka Görge, Anna K. Bernzen, Janna Clara Pirk and Maximilian Poretschkin*

**TUESDAY, OCTOBER 22**

- 294      Foregrounding Artist Opinions: A Survey Study on Transparency, Ownership, and Fairness in AI Generative Art  
*Juniper Lovato, Julia Zimmerman, Isabelle Smith, Peter Dodds and Jennifer Karson*
- 299      Annotator in the Loop: A Case Study of In-Depth Rater Engagement to Create a Prosocial Benchmark Dataset  
*Sonja Schmer-Galunder, Ruta Wheelock, Zaria Jalan, Alyssa Chvasta, Scott Friedman and Emily Saltz*
- 311      Outlier Detection Bias Busted: Understanding Sources of Algorithmic Bias through Data-centric Factors  
*Xueying Ding, Rui Xi and Leman Akoglu*
- 342      Representation Bias of Adolescents in AI: A Bilingual, Bicultural Study  
*Robert Wolfe, Aayushi Dangol, Bill Howe and Alexis Hiniker*
- 355      On The Stability of Moral Preferences: A Problem with Computational Elicitation Methods  
*Kyle Boerstler, Vijay Keswani, Lok Chan, Jana Schaich Borg, Vincent Conitzer, Hoda Heidari and Walter Sinnott-Armstrong*
- 399      As an AI Language Model, “Yes I Would Recommend Calling the Police”: Norm Inconsistency in LLM Decision-Making  
*Shomik Jain, D Calacci and Ashia Wilson*
- 419      Why Am I Still Seeing This: Measuring the Effectiveness of Ad Controls and Explanations in AI-Mediated Ad Targeting Systems  
*Jane Castleman and Aleksandra Korolova*
- 457      AIDE: Antithetical, Intent-based, and Diverse Example-Based Explanations  
*Ikhtiyor Nematov, Dimitris Sacharidis, Katja Hose and Tomer Sagi*
- 470      Estimating Weights of Reasons using Metaheuristics: A Hybrid Approach to Machine Ethics  
*Benoît Alcaraz, Aleks Knoks and David Streit*
- 490      Misrepresented Technological Solutions in Imagined Futures: The Origins and Dangers of AI Hype in the Research Community  
*Savannah Thais*

***Student Posters***

- 2      Creating an AI Data Context Index for Responsible AI in the Global South  
*Abiola Azeez*
- 40      One Bad NOFO? Federal Grantmaking Agency Silence on AI Governance  
*Dan Bateyko*
- 19      Schools of AI in the Public Sector: Fairness and Accountability Concerns  
*Marc Elliott*

- 37 Two-Stage Refugee Resettlement Models: Computational Aspects of the Second Stage  
*Simon Schierreich*
- 27 Decoding Global AI Governance: A Computational Linguistic Analysis of National Regulations  
*Eryclis Silva*
- 38 Bridging Ethics and AI: A Path to Moral Machines  
*Aisha Aijaz*
- 12 AI Ethics for Creativity  
*Alayt Abraham Issak*
- 10 The Main Challenges of AI Ethics:  
Historical Contextualization, Black-Boxing, Social Biases, Labor Invisibility  
*Konstantinos Konstantis*
- 5 Ethical Alignment in LLMs: What can metaethics tell us about LLM alignment?  
*William Rathje*
- 28 The Need For Inclusive NLP: Addressing Sociodemographic Bias and  
Enhancing Sociotechnical Systems through Interdisciplinary Frameworks  
*Pranav Narayanan Venkit*

9:00–10:45 **Poster Session 3 & Breakfast // Room LL20CD**

**Posters**

- 5 What to Trust When We Trust Artificial Intelligence  
*Duncan Purves, Schuyler Sturm and John Madock*
- 19 Are Large Language Models Moral Hypocrites? A study based on Moral Foundations  
*Jose Luiz Nunes, Guilherme Almeida, Marcelo Araujo and Simone Diniz Junqueira Barbosa*
- 78 Breaking the Global North Stereotype: A Global South-centric Benchmark Dataset for Auditing and Mitigating Biases in Facial Recognition Systems  
*Siddharth Jaiswal, Animesh Ganai, Abhisek Dash, Saptarshi Ghosh and Animesh Mukherjee*
- 83 Trusting Your AI Agent Emotionally and Cognitively: Development and Validation of a Semantic Differential Scale for AI Trust  
*Ruoxi Shang, Gary Hsieh and Chirag Shah*
- 113 Human vs. Machine: Behavioral Differences Between Expert Humans and Language Models in Wargame Simulations  
*Max Lamparth, Anthony Corso, Jacob Ganz, Oriana Skylar Mastro, Jacquelyn Schneider and Harold Trinkunas*
- 132 All Too Human: Understanding and Mitigating the Risk from Anthropomorphic AI  
*Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel and Verena Rieser*
- 139 Representation Magnitude has a Liability to Privacy Vulnerability  
*Xingli Fang and Jung-Eun Kim*
- 164 Afrofuturist Values for the Metaverse  
*Theresa Hice-Fromille and Sarah Papazoglakis*
- 191 Anticipating the risks and benefits of counterfactual world simulation models  
*Lara Kirfel, Rob MacCoun, Thomas Icard and Tobias Gerstenberg*
- 206 Surviving in Diverse Biases: Unbiased Dataset Acquisition in Online Data Market for Fair Model Training  
*Jiashi Gao, Ziwei Wang, Xiangyu Zhao, Xin Yao and Xuetao Wei*
- 214 Public vs Private Bodies: Who Should Run Which Advanced AI Audits and Evaluations? Evidence from Nine Case Studies of High-Risk Industries  
*Merlin Stein, Theresa Kriecherbauer, Amin Oueslati and Robert Trager*
- 229 Ecosystem Graphs: Documenting the Foundation Model Supply Chain  
*Rishi Bommasani, Dilara Soylu, Thomas Liao, Kathleen Creel and Percy Liang*
- 252 How Are LLMs Mitigating Stereotyping Harms? Learning from Search Engine Studies  
*Alina Leidinger and Richard Rogers*

- 277 Human-Centered AI Applications for Canada's Immigration Settlement Sector  
*Isar Nejadgholi, Maryam Molamohammadi, Kimiya Missaghi and Samir Bakhtawar*
- 295 Compute North vs. Compute South: The Uneven Possibilities of  
Compute-based AI Governance Around the Globe  
*Vili Lehdonvirta, Bóxi Wú and Zoe Hawkins*
- 322 Interpretations, Representations, and Stereotypes of Caste within Text-to-Image Generators  
*Sourojit Ghosh*
- 324 Ontology of Belief Diversity: A Community-Based Epistemological Approach  
*Richard Zhang, Erin Van Liemt and Tyler Fischella*
- 332 Algorithm-Assisted Decision Making and Racial Disparities in Housing:  
A Study of the Allegheny Housing Assessment Tool  
*Lingwei Cheng, Cameron Drayton, Alexandra Chouldechova and Rhema Vaithianathan*
- 340 On the Pros and Cons of Active Learning for Moral Preference Elicitation  
*Vijay Keswani, Vincent Conitzer, Hoda Heidari, Jana Schaich Borg and Walter Sinnott-Armstrong*
- 344 A Model- and Data-Agnostic Debiasing System for Achieving Equalized Odds  
*Thomas Pinkava, Jack McFarland and Afra Mashhadi*
- 347 Particip-AI: A Democratic Surveying Framework for  
Anticipating Future AI Use Cases, Harms and Benefits  
*Jimin Mun, Liwei Jiang, Jenny Liang, Inyoung Cheong, Nicole DeCairo, Yejin Choi,  
Tadayoshi Kohno and Maarten Sap*
- 360 Legitimizing Emotion Tracking Technologies in Driver Monitoring Systems  
*Aaron Doerfler and Luke Stark*
- 373 Automating Accountability Mechanisms in the Judicial System Using LLMs:  
Opportunities and Challenges  
*Ishana Shastri, Shomik Jain, Barbara Engelhardt and Ashia Wilson*
- 374 Do Responsible AI Artifacts Advance Stakeholder Goals?  
Four Key Barriers Perceived by Legal and Civil Stakeholders  
*Anna Kawakami, Jordan Taylor, Sarah Fox, Haiyi Zhu and Ken Holstein*
- 388 Dataset Scale and Societal Consistency Mediate Facial Impression Bias in Vision-Language AI  
*Robert Wolfe, Aayushi Dangol, Alexis Hiniker and Bill Howe*
- 392 Not Oracles of the Battlefield: Safety Considerations for AI-Based Military Decision Support Systems  
*Emilia Probasco, Matthew Burtell, Helen Toner and Tim G. J. Rudner*
- 413 Vernacularizing Taxonomies of Harm is Essential for Operationalizing Holistic AI Safety  
*Wm. Matthew Kennedy and Daniel Vargas Campos*
- 427 Measuring Human-AI Value Alignment in Large Language Models  
*Hakim Norhashim and Jungpil Hahn*

- 431 Sponsored is the New Organic: Implications of Sponsored Results on Quality of Search Results in the Amazon Marketplace  
*Abhisek Dash, Saptarshi Ghosh, Animesh Mukherjee, Abhijnan Chakraborty and Krishna P. Gummadi*
- 444 Navigating Governance Paradigms: A Cross-Regional Comparative Study of Generative AI Governance Processes & Principles  
*Jose Luna, Ivan Tan, Xiaofei Xie and Lingxiao Jiang*

**Student Posters**

- 32 Data Cleaning, Discard Studies, and Discretionary Power  
*Pinar Barlas*
- 7 Enhancing Transparency and Research Ethics through Human AI Techniques  
*Tatiana Chakravorti*
- 14 Model Multiplicity for Responsible AI  
*Prakhar Ganesh*
- 36 Automated Decision-Making Systems for Behavioral Regulation: Understanding Perceptions and Behavioral Reactions  
*Carmen Loefflad*
- 23 Intent-aware Example-based Explainability  
*Ikhtiyor Nematov*
- 35 Historiography of the Boundary-Works in Artificial Intelligence  
*Sokion Choi*
- 9 Enhancing Human-AI Collaboration through Adaptive Interaction and Explainability  
*Zhaobin Li*
- 11 Making Sense of Digital Domination  
*Jonne Maas*
- 30 "Computer says no": Impacts of AI on rural SMEs  
*Susan Sheldrick*
- 3 Quantitative and Organizational Approaches to Epistemic Risk in Generative and General-Purpose AI  
*Robert Wolfe*

**10:45–11:45 Oral Session 10 // Room LL20AB**

- 483 LLM Voting: Human Choices and AI Collective Decision-Making  
*Joshua C. Yang, Damian Dalisan, Marcin Korecki, Carina I. Hausladen and Dirk Helbing*
- 424 Breaking Bias, Building Bridges: Evaluation and Mitigation of Social Biases in LLMs via Contact Hypothesis  
*Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos and Ziwei Zhu*

# WEDNESDAY, OCTOBER 23

- 426      Understanding Intrinsic Socioeconomic Biases in Large Language Models  
*Mina Arzaghi, Florian Carichon and Golnoosh Farnadi*
- 400      Identifying Implicit Social Biases in Vision-Language Models  
*Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen and Marzyeh Ghassemi*
- 11:45–12:15      **Break**
- 12:15–1:15      **Keynote 3 // Room LL20AB**  
danah boyd, Microsoft Research / Georgetown University
- 1:15      **Closing Remarks // Room LL20AB**

# AIES-24 SPONSORS

AIES 2024 would like to thank the generous sponsors who allowed us to support Ph.D. students, invited speakers, social events, and to reduce the registration fee.

## PLATINUM



## SILVER



## BRONZE



### HOSTED BY



### GENEROUS FINANCIAL SUPPORT FOR THE STUDENT PROGRAM PROVIDED BY



## ORGANIZING ASSOCIATIONS

