THE PROGRAM GUIDE FOR THE

Ninth AAAI/ACM Conference on

ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

OCTOBER 12–14, 2026 | MALMÖ, SWEDEN

PROGRAM CONTENTS

Acknowledgements **3**

Welcome **4**

Program Overview **7**

Detailed Program **8**

Monday 8

Tuesday 16

Wednesday 25

Sponsors **32**

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 Ninth Annual AAAI / ACM Conference on Artificial Intelligence, Ethics, and Society.

CONFERENCE CO-CHAIRS

Virginia Dignum *Umeå University*
Petter Ericson *Umeå University*

CONFERENCE PROGRAM CO-CHAIRS

Aylin Caliskan *University of Washington*
S. Matthew Liao *New York University*
Mirko Schäfer *Utrech University* *(Active in the program until mid-June 2026)*
Daniel S. Schiff *Purdue University*

STUDENT PROGRAM CHAIRS

Katerina Drakos *University of Copenhagen*
Bertilla Fabris *Umeå University*
Tracie Farrell *The Open University*
Ángel Pavón Pérez *The Open University*

The AIES-26 Program Committee welcomes you to Malmö!

WELCOME FROM THE CONFERENCE CO-CHAIRS

The AIES-26 Conference Chairs welcome you to Malmö! This is the 9th edition of the AIES conference, which started from the premise that the ethics and societal impact of AI deserve the same scientific rigour as the systems themselves. A premise that is contested more openly now than ever, which makes our case not easier, but definitely more urgent!

Our goal in organising AIES is to bring together a community of scientists and researchers who care, both about their subject matter and about each other, about scientific rigour, about ethics, and about society. The continuing increase in submission volume, while very challenging for the Program Committee and its chairs, is a testament to the importance of the issues we discuss here, and to the increasing relevance of this conference. We hope that this year's AIES will be an opportunity to learn, organise, discuss, and explore the multitude of important and interesting issues facing us. We also hope that the conference serves as a way to reconnect with old friends and colleagues, continue conversations from before, and to start new things, connections, friendships, and projects, that will continue into the future.

Travel to Sweden is not the easiest nor the cheapest, and we acknowledge that many of you are here at real cost. All the more reason that we are very happy to see so many of you here. Sweden is often seen as a frontrunner in digitalization, and many actors here are highly ambitious when it comes to AI and its use across society. This makes Malmö a fitting place to ask the questions AIES is about: who benefits from AI systems, who carries their costs, and how we would know.

We wish you all a very fruitful AIES 2026!

Virginia Dignum, Petter Ericson, *Umeå University*

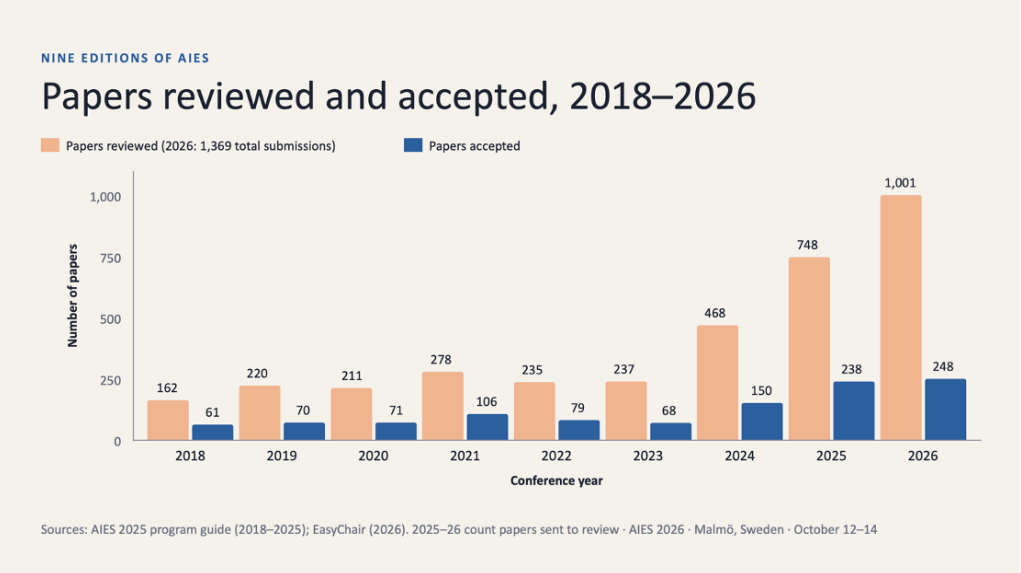
WELCOME FROM THE
CONFERENCE PROGRAM CO-CHAIRS

Welcome to the 9th annual AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society in Malmö! We are grateful to our hosts at Malmö Live, to the Conference chairs Virginia Dignum and Petter Ericson at Umeå University; the Student Co-chairs Katerina Drakos (University of Copenhagen), Bertilla Fabris (Umeå University), Tracie Farrell (Open University), and Ángel Pavón Pérez (Open University) and to Meredith Ellison, Briana Kemeny, Francisco Cruz, and the rest of the AAAI staff, whose work made it possible for us to concentrate on assembling the program.

When AIES first convened in 2018, it was a small gathering of 57 papers co-located with the main AAAI conference, and AI ethics was a niche concern. Eight years later, the AI systems we study are no longer experimental artifacts. They are infrastructure, shaping governance, labor, knowledge production, and daily life. The questions that brought this community together have moved to the center of public argument, and the pace of this shift has made our work more necessary than ever.

This year, we sent 1,001 papers out for review and accepted 249, again a record on both counts. We are deeply grateful to 99 Senior Program Committee Members, 940 Program Committee Members, 63 Student Track Program Committee Members, and 70 external reviewers from all over the world who volunteered their time to provide peer reviews for a submission pool of this size and disciplinary range.

The program this year has oral presentations or discussions of 44 papers and three poster sessions of 205 papers. No one of these is more prestigious than the others. Session assignments were made to build a coherent and engaging program within the constraints of three days, and we would encourage you to read the poster sessions as the intellectual core of the conference. We are also glad to welcome 25 students participating in the discussion panels and presenting their posters.



Continued on next page

AIES OVER TIME

From a co-located workshop-sized meeting to 1,000+ reviewed papers

Year	Location	Reviewed	Accepted	Rate	SPC + PC
2018	New Orleans, USA	162	61	38%	44
2019	Honolulu, USA	220	70	32%	88
2020	New York, USA	211	71	34%	80
2021	Virtual	278	106	38%	93
2022	Oxford, UK	235	79	34%	163
2023	Montréal, Canada	237	68	29%	139
2024	California, USA	468	150	32%	219
2025	Madrid, Spain	748	238	32%	690
2026	Malmö, Sweden	1,001	248	25%	≈1,040

Sources: AIES 2025 program guide; EasyChair; AIES websites. 2024 location: verify (Bay Area, CA) · AIES 2026 · Malmö, Sweden · October 12–14

We are pleased to have four distinguished keynote speakers: Sandra Wachter (University of Oxford), on what deregulation and the AI hype cycle cost human rights protection; Stefan Larsson (Lund University) on AI Governance in the Nordic Public Sector; Irina Shklovski (University of Copenhagen), on why responsible AI keeps falling short of its own promises; and Tadayoshi Kohno (Georgetown University) on Thoughts for the AI and AI Ethics Community, from a Computer Security Research Perspective. Our panel, Who Sets the Terms? AI Governance as Geopolitics, moderated by Virginia Dignum and opened by Kim van Sparrentak MEP, brings together Melody Musoni, Paul Verhagen, Jimena Viveros, and Matthias Spielkamp to ask what is actually being decided about AI, where, and by whom.

We hope you find the conference generative and rewarding, and that you leave with new friends and new ideas.

Aylin Caliskan *University of Washington* · **S. Matthew Liao** *New York University*
Mirko Schäfer *Utrecht University* · **Daniel S. Schiff** *Purdue University*

WELCOME FROM THE
STUDENT PROGRAM CHAIRS

Student research is a glimpse into the future — where one can see the direction of travel and the new lines of inquiry that are being opened by exceptional students from all over the world. Every year, the AIES Student Program aims to facilitate student participation by offering financial support, connection with our wider community of world class researchers, and 1:2:1 mentorship. We invite you to welcome students into our interdisciplinary community by engaging with student work during the student discussion panel and poster sessions, and learning more about the exciting new perspectives they are bringing to the subject of Artificial Intelligence, Ethics and Society.

We really appreciate those senior scholars who are serving as mentors to the students during the conference, and also wish to thank SIGAI and AAAI for their generous financial support.

Tracie Farrell *The Open University, UK* · **Ángel Pavón Pérez** *The Open University, UK* ·
Katerina Drakos *University of Copenhagen, Denmark* · **Bertilla Fabris** *Umeå University, Sweden*

PROGRAM OVERVIEW

	MONDAY, OCTOBER 12	TUESDAY, OCTOBER 13	WEDNESDAY, OCTOBER 14
8:00 AM	Registration		
8:15 AM			
8:30 AM	Opening Remarks	Keynote 2: Stefan Larsson	Keynote 3: Irina Shklovski
8:45 AM			
9:00 AM	Discussion Session 1	Student Panel Discussions	Discussion Session 4
9:15 AM			
9:30 AM			
9:45 AM			
10:00 AM			
10:15 AM	Poster Session 1 and Refreshment Break	Poster Session 2 and Refreshment Break	Poster Session 3 and Refreshment Break
10:30 AM			
10:45 AM	Paper Session 1	Paper Session 3	Paper Session 5
11:00 AM			
11:15 AM			
11:30 AM			
11:45 AM			
12:00 PM	Lunch	Lunch	Informal Town Hall
12:15 PM			Lunch
12:30 PM			
12:45 PM			
1:00 PM			
1:15 PM			
1:30 PM	Keynote 1: Sandra Watcher	Policy Panel: Intro: Kim van Sparrentak (MEP) Melody Musoni, Paul Verhagen, Matthias Spielkamp, Jimena Viveros	Keynote 4: Tadayoshi Kohno
1:45 PM			
2:00 PM	Discussion Session 2		Paper Session 6
2:15 PM			
2:30 PM		Paper Session 4	
2:45 PM			
3:00 PM			
3:15 PM	Poster Session 1 (cont.) and Refreshment Break	Poster Session 2 (cont.) and Refreshment Break	Poster Session 3 (cont.) and Refreshment Break
3:30 PM			
3:45 PM	Paper Session 2	Discussion Session 3	
4:00 PM			Closing Remarks and Awards
4:15 PM			
4:30 PM			
4:45 PM			
5:00 PM	Poster Session 1	Poster Session 2	
5:15 PM			
5:30 PM			
5:45 PM			

MONDAY, OCTOBER 12

8:00 AM **Registration**

8:30 AM **Opening Remarks**

9:00 AM **Discussion Session 1: Participation, Sovereignty, and Power in AI Governance**
Chair: Haiyi Zhu

258 The Mitigation Paradox: How AI-Disinformation Experts Converge on Threats but Polarize on Solutions
Alexander Loth, Mina Schütz, Martin Kappes and Marc-Oliver Pahl

1125 Appearing Legitimate is Not Enough: Interrogating Synthetic Agents in Representational Processes through a Participatory Design Lens
Aditya Nayak, Aditi Vashistha, Alissa Centivany and Aakash Gautam

1284 Cognitive Digital Twins: Ethical Risks and Governance for AI Systems That Model the Mind
Vamshi Krishna Bonagiri, Juan Nicolás Sepúlveda-Arias, Abdoul Jalil Djiberou Mahamadou and Monojit Choudhury

1313 Sovereignty Theater: Why Sovereign AI Falls Short of Cultural Self-Determination
Bruria Friedman Feldman, Tyler Smith, Victor Y. Wu and Yuanhao Zou

10:15 AM **Poster Session 1 and Refreshment Break**

10:45 AM **Paper Session 1: Bias and Discrimination in AI Systems**
Chair: Franziska Roesner

88 Fundamentally Unfair: Analysis of Gendered Discourse in U.S. Capital Trials
Camille Harris, Emma Harvey, Andrea W Wen-Yi, Sandra Babcock, Rachel Goldberg, Amanda Potts and Allison Koenecke

377 ABLEIST: Intersectional Disability Bias in LLM-Generated Hiring Scenarios
Mahika Phutane, Hayoung Jung, Matthew Kim, Tanu Mitra and Aditya Vashistha

1007 Examining Racial and Gender Bias in CLIP via Explainability-Based Visual Attention Analysis
Tesnim Gulsen and Pınar Uluer

1156 Linguistic Triggers of Gender and Racial Bias in Open-Weight LLMs Applied to Recruitment
Kosuke Kitahara and Nobuhiro Yamaguchi

1299 A Proxy-Based Method for Measuring Racial Disparities in Text-to-Image Generators
Nia Mucher, Ngan Pham, Jason Chin, Aaron Brown and David Evans

12:00 PM **Lunch Break (on your own)**

1:30 PM

Keynote 1: How geopolitics, deregulation, and the AI hype cycle threaten human rights protection

Prof. Dr. Sandra Wachter

Abstract: It is impossible to open a newspaper and not see yet another story about AI and its promise to enable economic growth. Nowadays it would almost border on heresy to stand in the way of AI development. This can clearly be seen in the current political climate which makes one thing very clear: we have entered a new age of deregulation.

In recent months, Big Tech successfully lobbied for stopping and revising the EU AI Act, changing the EUs General Data Protection Regulation (GDPR), and shelving the promised AI Liability Directive. "Cutting the red tape" is the declared goal in order to unleash the full potential of AI and enable competitiveness.

Interestingly enough, the counterfactual is never assessed. In other words, what are the costs of deregulation? How many people will now see no legal recourse if AI harms them? How many people's health will be compromised by weakened environmental laws? How many people's lives are now in danger due to the lower non-discrimination standards?

This begs the question: who actually benefits from deregulation?

This talk will discuss who really benefits from lowering legal frameworks. We will discuss Big Tech's AI promises, explore their current lobbying efforts and the costs society has to bare because of it.

However, I will show that profitability and protecting human rights do not have to be in conflict and will propose my new idea for tech governance. I will argue that the public is in all but name a shareholder of AI companies and thus is entitled to receive socially beneficial technology in return for its investment.

Bio: Prof. Dr. Sandra Wachter is Professor of Technology and Regulation at the University of Oxford at the Oxford Internet Institute and is Humboldt Professor of Technology and Regulation at the Hasso Plattner Institute. At the Oxford Internet Institute, Professor Sandra Wachter leads and coordinates the Governance of Emerging Technologies (GET) Research Programme that investigates the legal and ethical implications of AI, Big Data, and robotics as well as Internet and platform regulation. Professor Wachter also serves as a policy advisor for governments, companies, and NGOs around the world on regulatory and ethical questions concerning emerging technologies.

Professor Wachter publishes widely in top journals, including Science and Nature and focuses on legal, ethical, and technical aspects of AI and inferential analytics, explainable AI, algorithmic bias, platform regulation, profiling, as well as emotion- and facial-recognition software. The societal impact of generative AI and hallucinations in areas such as the future of work, misinformation, free press, and human rights is also at the heart of her research agenda.

Professor Wachter is an affiliate and member at numerous institutions, such as the Berkman Klein Center for Internet & Society at Harvard University, World Economic Forum's Global Futures Council on Values, Ethics and Innovation, UNESCO, European Parliament Working Group on AI Liability, Law Committee of the IEEE, World Bank's Task Force on Access to Justice and Technology, United Kingdom Police Ethics Guidance Group, British Standards Institution, Law Faculty at Oxford, Bonavero Institute of Human Rights, Oxford Martin School and the Oxford University Press.

Previously, Professor Wachter was a visiting Professor at Harvard Law School. Prior to joining the OII she studied at the University of Oxford and the Law Faculty at the University of Vienna. She has also worked at the Alan Turing Institute, Royal Academy of Engineering and the Austrian Ministry of Health.

Professor Wachter has been the subject of numerous media profiles, including by the Financial Times, Wired, Nature, TechCrunch, der Spiegel, and Business Insider. Her work has been prominently featured in several documentaries, including pieces by Wired , Reuters, and the BBC, and has been extensively covered by The New York Times, Time Magazine, Reuters, Forbes, Fortune, CNN, Harvard Business Review, Guardian, BBC, Telegraph, CNBC, CBC, Huffington Post, Washington Post, Science, Nature, MIT Tech Review, New Scientist, HBO, The Sunday Times, and Vice Magazine.

Professor Wachter has received numerous awards, including the Alexander von Humboldt Foundation Research Award (2025) which grants EUR €3,5 Mio in funding, the O2RB Excellence in Impact Award (2018 and 2021), the Computer Weekly Women in UK Tech award (2021), the Privacy Law Scholar (PLSC) Award (2019) for her paper A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI, and the CognitionX Outstanding Achievements & Research Contributions – AI Ethics Award X (2017 and 2023) for her contributions in AI governance.

2:00 PM **Discussion Session 2: Trust, Dependence, and Human Oversight**
Chair: Ken Holstein

128 Mediational Opacity: How AI Reshapes Our Epistemic Environments, and How to Audit It
Alessandro Facchini

766 The Perils of Agency: How Developers Perceive, Prioritize, and Address Risks in Agentic AI Products
Hao-Ping Lee, Jessica He, David Piorkowski, Thomas Serban von Davier, Jodi Forlizzi and Sauvik Das

1013 "Always Want to Use It for Everything": Understanding Young Adults' Perceptions of AI Dependence
Ashlee Milton, Leah Ajmani, Amy Heger, Forough Poursabzi-Sangdeh, Mihaela Vorvoreanu and Jina Suh

1282 Epistemic Trustworthiness in Generative AI: A Normative Framework for Warranted Reliance in High-Stakes Workflows
Nimisha Karnatak, Max Van Kleek and Nigel Shadbolt

3:15 PM **Poster Session 1 (cont.) and Refreshment Break**

3:45 PM **Paper Session 2: AI Across Work and Education**
Chair: Miriam Fernandez

832 Who Will Become the Next Senior? How Generative AI Erodes the Development Pathway in Software Engineering
Sumin Yu and Taesup Moon

887 Cost-of-Ethics Crisis: Beliefs, Decisions, and Justifications in the Job Searches of Computer Science Students in Canada and the United States
Mohamed Abdalla, Sahar Abdalla, Alicia Cappello, Kyrie Dowling, Danaé Metaxa, David Widder and Catherine Stinson

901 Beyond Language Resource Tier: A Survey Study of Culturally Situated Resource Scarcity in LLM-Generated Cultural Translation of Math Word Problems
Parisa Suchdev and Juniper Lovato

1020 On the Generalizability of Fair Representations
Shai Ben-David, Tosca Lechner and Ruth Urner

1401 An Agent-Based Simulation of Fairness Interventions in Hiring
Mattia Cerrato, Herbert Dawid and Stefan Kramer

MONDAY, OCTOBER 12

5:00 PM **Poster Session 1**

Posters

39 The Unfalsifiable Delusion: Mass Surveillance Infrastructure and Disparate Impact on Americans with Paranoia-Spectrum Conditions
Travis Gilly

41 LLMs on the Political Compass: Prompt-Conditioned Positioning and Directional Drift
Rim Ben Salem, Antonio Vincent Romeo and Esma Aïmeur

73 From Oracles to Productive Friction: Designing for Meaningful Human Oversight in Clinical AI
Elisa Rubegni, Chiara Natali, Clara Valsecchi, Stefania Rizzo, Omran Ayoub and Alessandro Facchini

149 When AI Becomes Routine: A Decade of Public AI Mediation in Korean Go Commentary
Haewoon Kwak

169 Explaining Time Series Classifications Counterfactually: A Human-Centered Evaluation
Eoin Delaney, Greta Warren, Derek Greene and Mark T. Keane

194 CompanionSim: Synthetic Data for Evaluating Anthropomorphism in Human-AI Relationships
Jacy Anthis, Mark Diaz and Renee Shelby

236 Engineering Discipline to Governance Economics: A Proportionate Two-Pass Framework for Algorithmic Monitoring in Capital Markets
Anchal Gupta, Anupam Jain, Gleb Papyshev and James T Kwok

252 From Model Approval to Interface Governance: Federated Learning, Lifecycle Accountability, and AI-Enabled Software as a Medical Device
Liu Ben

264 What Should I Become? When LLMs Present a Slice of Opportunity as the Whole
Taylor Wise, Scott Seidenberger and Anindya Maiti

283 What Does the Credential Still Certify? Cognitive Stewardship for AI-Mediated Education
Kai Yao

288 Speech in the Panaudikon: On the Relationship Between Speech Processing Research and Surveillance
Vasilisa Rudykh, Zeerak Talat and Peter Bell

300 When Reasoning Hurts Confidence: Demographic Calibration of Vision-Language Models in Medicine (Extended Abstract)
Sonnet Xu and Roxana Daneshjou

307 The Alignment Problem Is an Identity Problem: Lessons from Four Billion Years of Major Evolutionary Transitions
Nyx Redondo

309 When AI Does the Work, What Is Learning For? Post-Instrumental Learning and the Risk of Capacity Dissolution
Kai Yao

332 Limits of Interventions in AI-Assisted Decision Making
Shikha Shrestha, Eunhye Ahn and Yevgeniy Vorobeychik

MONDAY, OCTOBER 12

346 How Are Emerging Adults Using ChatGPT in the Wild?
Marx Wang, Jade Li, Songling Ngo, Raghavi Putluri, Rachel Hanebutt, Joy Chang, Robert Wolfe, Laura Boccia, Oluwaferanmi Omitoyin, Anind Dey, Katie Davis and Alexis Hiniker

361 Governing AI by Standards: Mapping AI Governance Themes in ISO/IEC and National AI Standards
Marta Lasek-Markey, Delaram Golpayegani, Aphra Kerr and Dave Lewis

374 Bridging Interpretation and Automation: A Bibliometric Review of AI-Assisted Qualitative Research (Extended Abstract)
Dong-Hyeon Im, Cheongil Kim, Ji-Bum Chung, Seunghoo Jeong and Byeong-Je Kim

389 Generative AI and Academic Self-Confidence: Understanding How AI Tools Influence English Learning Motivation Among Multilingual and International Students
Sehrish Daud and Sourojit Ghosh

401 Sources of Inequity and Fairness Risks in Wellbeing Sensing
Han Zhang, Vedant Das Swain, Koustuv Saha, Anind Dey and Jennifer Mankoff

416 Measuring and Mitigating (Over)reliance to Build Human-Compatible AI
Lujain Ibrahim, Katherine M. Collins, Sunnie S. Y. Kim, Anka Reuel, Max Lamparth, Kevin Feng, Lama Ahmad, Prajna Soni, Alia El Kattan, Merlin Stein, Siddharth Swaroop, Vishakh Padmakumar, Ilia Sucholutsky, Andrew Strait, Diyi Yang, Q. Vera Liao and Umang Bhatt

439 Context and Symmetry in Auditing: A Case Study of Skeleton Inference in Motion Capture
Emma Harvey, Emanuel Moss, Hauke Sandhaus, Abigail Z. Jacobs and Mona Sloane

450 Algorithmic Shortlisting in Participatory Budgeting
Juan Ignacio Zambrano, Clément Contet, Jairo Gudiño-Rosero, Felipe Garrido-Lucero, Umberto Grandi and César A. Hidalgo

462 Cross-Lingual Robustness of Societal Bias Representations in Large Language Models: A Contrastive Activation Steering Analysis of Llama 2 and GPT-OSS
Mehuly Chakraborthy and Antonio Rodà

477 Human Rights by Design for AI Technologies: Towards Outlining and Operationalizing an Emerging Field of Research and Practice
Leonardo Iwaya, Samuel Wairimu, Tara Mehrabi, Annika Rudman and Simone Fischer-Hübner

491 Qualifying and Quantifying Risk Under the EU AI Act
Gustavo Gil Gasiola, Sarah Cen and Frederike Zufall

507 When Distrust in Humans Drives Trust in AI: Institutional Bias and Algorithmic Preference Among Marginalized Groups (Extended Abstract)
S Mo Jones-Jang, Rylee Ahn and Jiyoung Lee

512 AI Localism - Counter Imaginaries from City AI Governance
Bhavana Bheem, Samuel Gerstein, Karrie Karahalios and Madelyn Rose Sanfilippo

523 Why Do AI Agents Break Rules? How Framing, Context, and Social Signals Shape Compliance
Mika Okamoto, Ansel Kaplan Erol and Kutluhan Erol

540 Calibrating WEAT Against Anisotropy: ZCA Whitening as a Geometric Pre-Processing Step for Embedding Association Tests
Seitaro Ono, Senna Ross and Jun Saiki

MONDAY, OCTOBER 12

550 We Need Explanation Cards to Connect Explanation Algorithms to the Real World
Eric Günther, Balázs Szabados, Kristof Meding, Gunnar König, Sebastian Bordt and Ulrike von Luxburg

566 A Method for Learning Value Systems in Generative AI
Andrés Holgado-Sánchez, Holger Billhardt and Sascha Ossowski

580 Spotting (and Missing) Algorithmic Bias: Investigating User Understanding in a Fairness Assessment Tool
Anna Verheyden, Yizhe Zhang, Robin De Croon, Simone Stumpf and Katrien Verbert

619 Centering Civil Society Voices for Social Justice: Critical Reflections and Ways Forward for AI Practitioners
Steven Vethman, Marie Mirsch, Louise Aupetit and Emma Kallina

629 Learning by Chatting? Investigating the Impact of Generative AI on Information Seeking and Learning
Shravika Mittal, Su Lin Blodgett and Q. Vera Liao

642 "How Much Do You Want to Play God in Controlling the Dataset?": Data Scientists' Perspectives on Modifying Data to Address Fairness Concerns
Kevin Bryson, Galen Harrison, Margaret Jennings, Alex Kale and Blase Ur

651 The Forager's Dilemma: Peer Enforcement Alone Is Not Enough for Stable Multi-Agent Alignment
Majid Ghasemi, Prabhjyot Singh and Mark Crowley

667 Visions of Disciplinary and Education Futures with Generative AI
Jaemarie Solyst, Amy J. Ko and R. Benjamin Shapiro

682 Auditing Alignment Controllability in LLMs via Political Axes
Bartol Bućan, Nikola So·ec, Sarah Isufi, Morena Granić, Luka Hobor, Agneza Krajna, Mihael Kovac and Mario Brcic

702 Diet: Interpreting Ethical Tradeoffs in Disrupted Sociotechnical Systems
Divya Sundaresan, Gordon Briggs and Munindar P. Singh

723 "Death By A Thousand Taxonomies?": AI Risk Classification In Practice
Glen Berman, Ned Cooper, Angel Hsing-Chi Hwang, Wesley Hanwen Deng, Renee Shelby and Ben Hutchinson

735 What Should AI Generate? Moving Beyond Bias in Generative Systems
Jennifer Chien

765 AI Assistance for Human Review of Default Judgments
Theodora Worledge, Othman Bensouda Koraichi, Daniel Bernal, Aviv Caspi, Tatsunori Hashimoto, Carlos Guestrin and David Freeman Engstrom

781 Beneficiary-Based Responsibility for AI Training Labour: Lessons from Kenya's Gig Economy and Beyond
Ruoyi Lu and Brenda Maina

789 Beyond Copy-Paste Regulation: Innovation-Weighted Risk Classification for AI Governance in Emerging Economies
Egemen Bora Tunçarslan, Ece Melisa Şerik and Samed Çayır

796 Follow the Norm: Accounting for Fine-Tuning and Prompt Effects on Model Rationales
Long Hoang Nguyen, Brice Valentin Kok-Shun, Guangyu Du and Ali Sunyaev

MONDAY, OCTOBER 12

824 From Compliance to Contextual Legitimacy: A Sociotechnical Framework for Evaluating AI Systems
Giuditta Bellosi and Roberta Calegari

845 A Crack in the Framework: Risk Management and the Governance of General-Purpose AI with Systemic Risks
Toni Lorente

886 Fairness Detectability: Do Explanations Recover Fairness-Relevant Causal Effects?
Sakina Hansen, Kateryna Koval, Anna-Alina Bondarets and Joshua R. Loftus

902 A Precarious Normative Balance or the "Third Way"? Sociotechnical Imaginaries in India's AI Governance
Deepshikha Sharma, Suriya Krishna B S, Manohar Kumar and Praveen Priyadarshi

923 Small-world Networks of Agents Brainstorm AI Risks to Support Ideation
Ke Zhou, Edyta Bogucka and Daniele Quercia

950 Bridging AI Research and Policy: A Co-Production Framework and the Case for Policy Transfer Offices
Malihe Alikhani, Jeongeun Kim, Anna Haensch, Saki Imai, Serena Booth, Andrea Stith and Adriana Bankston

961 The Artificial Moral Machine: Do LLMs Mirror Human Moral Diversity or Merely Caricature It?
Davide Liga, Liuwen Yu, Marija Slavkovik, Samia Touileb, Beishui Liao and Leendert van der Torre

984 Do Personalized Evaluation Functions Reflect Human Preferences? A Study of Weighted Proximity in Algorithmic Recourse
Firdaus Ahmed Choudhury, Ethan Leicht, Jude Ethan Bislig, Arsh Gupta, Akshaya Shyamsundar Rekha, Shubham Gupta, Haren Anand, Hangzhi Guo and Amulya Yadav

986 Applications of Risk Science to AI Fairness Evaluation: Principles, Challenges, and Best Practices
Kyra Wilson, Sabrina Kang, Saloni Dash and Aylin Caliskan

988 L2-Bench: An Evaluation Benchmark for Measuring LLM Capabilities in Second Language Education (Extended Abstract)
James Edgell, Wm. Matthew Kennedy, Ben Knight, Danielle Carvalho, Martin Ku and Isaac Pattis

1008 Wanted Dead or Alive: Anticipatory Regulation of Digital Immortality Platforms
Alexandra Gillespie, James Garforth and Stacy A. Doore

1022 Locating Translation as a Craft in the Age of AI
Daniel Chechelnitsky, Sireesh Gururaja, Seyi Olojo, Giuseppe Attanasio, Wesley Hanwen Deng, Chrysoula Zerva and Maarten Sap

1039 Whose "Intelligence" Are We Building? On Exclusion in Conceptual Frameworks of Artificial Intelligence
Dafna Burema and Vito Mengers

1064 Co-Constructing Sociotechnical AI Governance: Participatory System Mapping Using Algorithm Registers (Extended Abstract)
Íñigo de Troya, Maurus Enbergs, Neelke Doorn and Roel Dobbe

1068 Design Theater: Evaluating the Gap Between User-Facing Design Reasoning and Implementation in Generative UI Tools
Kashif Imteyaz, Kaif Imteyaz, Nakul Rajpal, Kaif Shaikh, Michael Muller and Saiph Savage

MONDAY, OCTOBER 12

1072	From Fair Representation to Just Recognition in Generative AI *Severin Engelmann and Daniel Susser*
1083	The Hidden Cost of a Friendly Face: Investigating the Effect of Agent Embodiment on Trust Loss in Conversational Recommender Systems After Misleading Health Claims *Haya Halimeh*
1086	Training on Mass Surveillance Datasets Provides No Empirical Benefits *A Pranav and Anne Lauscher*
1102	When Cooperation Predictions Become Reasons to Defect: Policy Defaults in LLM Assistance for Social Dilemmas *Kazunori Terada, Henrique Correia da Fonseca, Celso M. de Melo, Giorgio Coricelli and Jonathan Gratch*
1126	Asymmetric Trust Effects of Corrective AI in Expert Advisory Workflows Under Epistemic Dependence *Dennis Kim, Roya Daneshi, Bruce Draper and Sarath Sreedharan*
1174	Who Can Undo the Agent? Redressability as a First-Class Safety Property for Tool-Using AI Agents *Saad Alqithami and Abdulkareem Alzahrani*
1187	Ordinary, Reasonable Chatbots: Do AI Models Track Human Legal Judgments? *Nirav Patel, Emily Wenger and Christopher Buccafusco*
1225	The Politics of AI Ethics: How Normative Principles Translate into Law and Policy in South Korea *Chee Hae Chung and Daniel Schiff*
1245	Do Judges Behave like Algorithms? *Riya Manchanda, Eric Chen, Chloe Zhu, Cynthia Rudin, Brandon Garrett and Songman Kang*
1291	"It's OK Because...": Students' Rationalization of AI Use in Academic Writing *Jiyoon Kim, Kentaro Toyama, Sangmi Kim and John M. Carroll*
1328	Moral Missions: Sustaining Responsible Practice by Surfacing Practitioner Moral Decision-Making Procedures *Teanna Barrett, B. Biira, Jainaba Jawara, Andrew Shaw, Ziwei Dong, Chinasa T. Okolo, Seyi Olojo, Keerthana Kompella, Khadija Saho, Amy X. Zhang and Leilani Battle*
1395	User Perceptions of Responsibility and Autonomy of AI Consent Behaviours *Anastasia S. Apeiron, Pradeep K. Murukannaiah and Pinar Yolum*
1413	The Failure Happens Before the Drift: The Social Dynamics of Values in LLM Agent Societies *Farah Atif, Sougata Saha and Monojit Choudhury*
1444	Investigating Feature Substitution in Algorithmic Worker-Scoring Models *Abdulrahman Ababneh*
1466	The Promises and Limitations of Evaluation-Based AI Governance: Evidence from Expert Interviews *Lily Stelling, Jacob Charnock, Kevin L. Wei and David Krueger*

TUESDAY, OCTOBER 13

8:30 AM **Keynote 2: Chasing the Moving Target: AI Governance in the Nordic Public Sector**

Stefan Larsson

Abstract: Drawing from a concluding research project on AI governance in the public sector and an initiating project on copyright conflicts in AI model development, Stefan Larsson highlights regulatory challenges in AI. The Automated Administration (ADM-GOV) is a Finnish-Swedish project that has studied the evolving governance and regulation of AI and automated decision-making (ADM) in the public sector for four years. The insights are summarised and distributed in a print magazine, alongside various scientific papers. The magazine's creation was also an exploration of how AI tools impact scientific texts when summarised and the authors' experiences of this process. This provided insight into informal and formal norms surrounding AI-impacted media generation, which are in motion.

Bio: Stefan Larsson is an Associate Professor in Technology and Social Change at Lund University (LU), Sweden. As a lawyer and socio-legal researcher with PhDs in both social (Sociology of Law) and technical (Spatial Planning) sciences he leads an interdisciplinary research group on AI and Society at the faculty of engineering at LU. The group focuses on social and normative implications of AI and adaptive technologies in both private and public domains, ranging from how AI and automation affects or impacts the public sector and its governance to AI and prediction in health care, and social robotics.

9:00 AM **Student Panel Discussions**

1 Generated Data, Real Consequences: Why LLMs Are Inappropriate for Development Aid Targeting
Merritt Smith

21 Protecting the Mind: Privacy and Data Protection Challenges in Consumer Neurotechnology
Irene Rocchesso

28 Corporate Narratives of Care: Exploring Sociotechnical Imaginaries of AI in Swedish Eldercare
Lucy McCarren

29 Advancing Deliberation with AI: Architecture, Evaluation, and Simulation
Maurice Flechtner

37 AIES'26 Student Track Application: Pragmatic Explainable Machine Learning
Sakina Hansen

39 Toward Responsible and Sustainable AI Training Data Pipelines
Sophia Wilson

40 Do we need Computing Power Ethics? Participation Conditions in the AI Era
Yunqi Jiang

41 Towards Trustworthy Intelligent Tutoring Systems for Nursing Education: Integrating Nurse-Informed Design with a Frugal Neuro-Symbolic Architecture
Nevena Dicheva

TUESDAY, OCTOBER 13

42	Care-Full Alignment: Operationalizing Care Ethics for Human-LLM Relationships *Rayane El Masri*
44	Virtue by Observation: Social Reinforcement Learning for Intrinsically Ethical Agents *Majid Ghasemi*
55	Towards Culturally Grounded Emotional AI via a Community-Centred Approach *Himashi Rathnayake*
58	Ethical AI Tools for Critical Museum Practice *Erin Canning*
67	Distributed Responsibility, Contested Accountability: Auditing and Governing AI Systems Across the Supply Chain *Lelia Hanslik*
70	Community-Centered AI Governance *Bhavana Bheem*
72	Tracing Borrowed Authority in Relational AI Systems *Hanjing Shi*
75	The Prompt Engineering Fallacy: How Diverse Users Construct Prompts in Health Chatbot Interactions *Jia Lin*
76	Does the Brussels Effect Extend from Data Protection to AI Governance in Africa? Examining the Relationship Between European Frameworks and African Policy Adoption *Amaka Onebunne*
82	From Visibility to Accountability: Designing Actionable Transparency Infrastruc-tures for Public-Sector AI *Ana Gagua*
85	Integrating Computational and Community-Engaged Modeling to Advance Health Equity in Healthcare AI *Neha Srivathsa*
98	Structure, Meaning, and the Answerability of Automated Systems *Mattias Brännström*
100	Resisting Surveillance AI: Mechanisms of Accountability Evasion and Community Engagement *A Pranav*
101	Rethinking Agentic AI: A Sociotechnical Framework for Agents in Emergent Social Contexts *Marcelina Przespolewska*
105	Fair and Representative Clustering for Data Summarization *Ashwin Singh*
109	Inherent Politics of AI as a Matter of Justice (Student Program) *Laila Wegner*
120	Safe Enough to Try GenAI: A Multi-Level Sociotechnical Approach for Responsible Generative AI in Public-Serving Organizations *Ashley Khor*

10:15 AM **Poster Session 2 and Refreshment Break**

TUESDAY, OCTOBER 13

10:45 AM **Paper Session 3: Law, Regulation, and AI Accountability**

Chair: Suresh Venkatasubramanian

211 Governing by Proxy: How Regulatory Intermediaries Shape EU AI Act Compliance in Practice
Simon Jarvers and Orestis Papakyriakopoulos

905 Scaling, Lock-In, and Proxy Compliance: A Political Economy of Responsible AI
Florian A D Burnat and Brittany I Davidson

1087 Triangulating Across U.S. Federal AI Transparency Regimes
Emma Lurie, Emma Fauser, Qing He, Danaé Metaxa and Sorelle A. Friedler

1096 No One to Blame: A Framework of Constitutive AI Unaccountability
Long Hoang Nguyen, Eva Späthe, Sebastian Lins and Ali Sunyaev

1206 Govern, Map, Measure, Absorb: The Legibility Trap in Public Sector Participatory AI Governance
Srijan Pandey and Devansh Saxena

12:00 PM **Lunch Break (on your own)**

1:30 PM **Panel: Who Sets the Terms? AI Governance as Geopolitics**

Summary: Sovereignty is now the most used word in AI policy and the least supported by anything material. The AI Act made Europe the first jurisdiction to legislate comprehensively, and it is now cited in every direction: by American deregulators as a cautionary tale, by other governments as a template they lack the capacity to implement, by some European industry as the sole reason it cannot compete, and by European institutions as evidence of leadership. All there readings can be true at once, which is the problem. Most states govern systems they cannot inspect, running on compute they do not own, trained on data they never saw. This panel asks what is actually being decided, where, and by whom, and whether the answer is compatible with anything we would recognise as democratic authority.

Moderator: Prof. Dr. Virginia Dignum, Wallenberg Chair, Professor Responsible Artificial Intelligence and Director, AI Policy Lab AI Policy Lab | Department of Computing Science at Umeå University, Sweden

Bio: Virginia Dignum is Professor of Responsible Artificial Intelligence at Umeå University, Sweden, where she leads the AI Policy Lab. She is also senior advisor on AI policy to the Wallenberg Foundations and chair of the ACM's Technology Policy Council. She has a PHD in Artificial Intelligence from Utrecht University (2004), was appointed Wallenberg Scholar in 2024, is member of the Royal Swedish Academy of Engineering Sciences (IVA), and a Fellow of the European Artificial Intelligence Association (EURAI), and of ELLIS (European Laboratory of Learning and Intelligent Systems). She is also co-chair of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2.0, member of the UNESCO's expert group on the implementation of AI recommendations, the OECD's Expert group on AI, and founder of ALLAI, the Dutch AI Alliance. She has been a member of the United Nations Advisory Body on AI, the EU's High Level Expert Group on Artificial Intelligence, co-chair of the WEF's Global Future Council on AI, and leader of UNICEF's guidance for AI and children. In 2026, she received the Nordic DAIR Awards Lifetime Achievement in AI. She has over 300 peer-reviewed articles in major international journals, conferences and workshops with in total over 30000 citations. Her new book "The AI Paradox" (2026) is published by Princeton University Press.

TUESDAY, OCTOBER 13

Introductory Keynote: Kim van Sparrentak,Member of the European Parliament on behalf of the Groenlinks-PvdA delegation in the Greens/EFA

Bio: Kim van Sparrentak is a Member of the European Parliament on behalf of the Groenlinks-PvdA delegation in the Greens/EFA. She is Co-Chair of the LGBTIQ+ Intergroup, a member of the Committee on Internal Market and Consumer Protection, as well as a substitute in the Committee on Employment and Social Affairs. Kim's focus is on the human element in digitalisation, as well as digital sovereignty. She was negotiator for the Greens group on the AI Act, the Platform Work Directive and Parliament's lead negotiator on the Short-Term Rental Regulation and the own-initiative report on Addictive Design.

Panelist: Dr. Melody Musoni, Senior Digital Policy Analyst at ECDPM (the European Centre for Development Policy Management)

Bio: Melody Musoni is a senior digital policy analyst at ECDPM, where she leads research projects on the African digital economy. She holds a PhD in Law (Cybercrime), a Master's Degree in CyberLaw, a postgraduate diploma in Environmental Law, and an LLB degree from the University of Witwatersrand. She has published several academic papers and book chapters on topics including digital sovereignty, digital IDs, digital public infrastructure (DPI), artificial intelligence governance, and data protection. She actively engages with African and European policymakers on data and AI governance and previously assisted the SADC Secretariat with its data protection and compliance programs. She has over a decade of experience in legal practice, specializing in ICT law, data protection, and information security.

Panelist: Paul Verhagen, Senior Fellow, Geopolitics and Technology at The Hague Centre for Strategic Studies

Bio: Paul Verhagen (1991) is a Dutch-American-Chinese Senior Fellow at The Hague Centre for Strategic Studies on technology and geopolitics . He is the author of The Techtonic Shift, which argues that technological change has rendered our political institutions anachronistic, and that the return of geopolitics leaves little time to rebuild them. He comments regularly in Dutch and international media. His background is interdisciplinary, spanning theoretical physics, philosophy, artificial intelligence, and geopolitics.

Panelist: Jimena Sofia Viveros Alvarez, Founder and CEO of IQuilibriumAI and President of The HumAIne Foundation

Bio: Jimena Sofia Viveros Alvarez is a senior international lawyer, AI governance expert practitioner whose work bridges international security, international humanitarian law, and emerging technology policy. She is the Founder and CEO of IQuilibriumAI and President of The HumAIne Foundation, and was appointed as a Member of the United Nations Secretary-General's High-Level Advisory Body on Artificial Intelligence and as a GC-REAIM Commissioner, serving as Co-Lead for Peace and Security in both collegial bodies, Deputy Lead on the International Law Association Committee for AI and Technology Law, and Senior AI Expert at the OECD, UNESCO, ITU, ICRAC, CyberPeace Institute, among others. She served the Mexican Government in senior leadership positions as Director General and Head of International Affairs at the Ministry of Security, Tax and Finance Federal Prosecutors' Office, and Head Legal Advisor and Chief of Staff at the Federal Judicial Council and the Supreme Court. Her work focuses on the governance and accountability gaps generated by dual-use AI technologies across the military, law enforcement, and civilian domains, with particular attention to Global South deployments and to building inclusive, operationalizable multilateral frameworks. She has facilitated multistakeholder dialogue across governments, defense establishments, civil society, and industry, and is a widely published author and has presented at over 100 major international fora, and awaits her doctoral dissertation defense on dual-use AI governance this year.

Panelist: Matthias Spielkamp, Founder and Executive Director of AlgorithmWatch

Bio: Matthias Spielkamp is founder and executive director of AlgorithmWatch, a civil society organisation based in Berlin and Zurich. It advocates for algorithmic systems and AI that make societies more just, democratic, inclusive and sustainable. Matthias is a member to the European Commission's AI Act Advisory Forum and testified before committees of the Council of Europe, the European Parliament, the German Bundestag and other institutions on automation and AI. He holds master's degrees in Journalism from the University of Colorado in Boulder and in Philosophy from the Free University of Berlin.

2:30 PM Paper Session 4: AI in Health and Care

Chair: Dan Lizotte

492 Beyond "I Can't Help with That": How Child Safety Experts Evaluate AI Chatbot Safety
Hannah Cha, Neha Shukla, Solon Barocas, Alexandra Chouldechova, Eugenia Kim and Jennifer Wortman Vaughan

772 Nationwide AI Models for Depression Prediction: Regional Generalizability and Culturally Sensitive Variable Selection
Alissa Valentine, Helen Coupland, Eike Petersen, Emily Beaman, Merete Osler, Aasa Feragen and Melanie Ganz

1233 Same Question, Different Answers: Auditing Persona-Dependent Variation in Six LLMs on Smoking-Cessation Advice
Artin Yousefi, Aizierjiang Aiersilan, Lorien Abroms and David Broniatowski

TUESDAY, OCTOBER 13

3:15 PM **Poster Session 2 (cont.) and Refreshment Break**

3:45 PM **Discussion Session 3: AI at Work: Power and Social Justice**

Chair: Danaé Metaxa

500 Responsible AI as Hegemony: Epistemic Exclusion, Technosolutionist Conformity, and Performative Compliance Among AI Practitioners in India
Atmadeep Ghoshal, Arshia Dutta, Ruba Abu-Salma, Sanjay Modgil and Martim Brandao

578 "I Want to Be Pushed, I Want to Grow": Enabling Social Workers to Design Evaluations of LLM Augmentation in Their Work (Extended Abstract)
Anna Kawakami, Chloe Qianhui Zhao, Renee Shelby, Fernando Diaz, Haiyi Zhu and Ken Holstein

754 Unaccountable Delegation, Fading Skills: Mapping the Risks of Workplace AI Agents
Gabriele La Malfa, Lakmal Meegahapola, Edyta Bogucka, Jie M. Zhang, Michael Luck, Elizabeth Black and Daniele Quercia

1155 See Me, Believe Me: Causality, Intersectionality, and Interventions Improving the Appearance of Patients
Kenya Andrews, Mesrob Ohannessian and Elena Zheleva

5:00 PM **Poster Session 2**

Posters

43 AI Literacy: An Exercise in Power-Knowledge
Brady D. Lund and Zoe Abbie Teel

76 Hands-Off or Hands-On? Variation in Area Chair Practices and Implications for AI Support
Ines Arous, Neha Nayak Kennard, Andrei Mircea, Emily Kuang, Jackie Chi Kit Cheung and Andrew McCallum

85 Beyond Detection: Governing GenAI in Academic Peer Review as a Sociotechnical Challenge
Tatiana Chakravorti, Pranav Narayanan Venkit, Sourojit Ghosh and Sarah Rajtmajer

99 From Prediction Markets To Capability Markets: Epistemic Extractivism In Generative AI
Glen Berman and James Smithies

166 Safer On-Premise Medical AI: A Hybrid Framework for Hallucination Detection in Local LLMs
Sanne Wielinga, Clara Maathuis and Natasha Alechina

170 How to Catalyze Policy Guardrails against AI Chatbots' Potential Social-Emotional Harm: Lessons from Analogous Cases (Extended Abstract)
Ned Cooper, Ibrahim Emara and Qian Yang

200 "I Thought You Were the Uncensored Place": Norms, Rules, and Moderation in AI-Generated Sexual Content Communities
Lucy Qin, Jaron Mink and Elissa M. Redmiles

238 Explainable AI Isn't Enough! Rethinking Algorithmic Contestability
Timo Freiesleben, Kristof Meding and Gunnar König

TUESDAY, OCTOBER 13

247 Participatory Moral AI Is Not Neutral: The Invisible Hand of Developers
Taenyun Kim, Edyta Bogucka and Daniele Quercia

269 Hidden Errors in Big Data: The Case of Property Records
Evelyn Smith, Emma Harvey, Jacob S. Goldin and Daniel E. Ho

292 Designing for the Intentional Stance: Stance Misalignment, Value Alignment, and the Case for Care Ethics in Human-LLM Relationships (Extended Abstract)
Rayane El Masri, Aaron Snoswell and Ehsan Dehghan

303 The Epistemic Politics of AI Anthropomorphism
Donna M Bye and Levin Kuhlmann

863 Untangling Trust, Transparency, and Uncertainty in AI Literature and Policy
Kai Nylund, Shani Spivak, Dakuo Wang and Lace Padilla

443 From Values to Benchmarks: Evaluating Large Language Models for Governmental Use in Dutch
Laurens Samson, Iva Gornishka, Gossa Lô, Yuki M. Asano and Sennay Ghebreab

460 Law of Large Numbers: Accuracy as Statistical Measure for Compliance and Competition
Rabanus Derr, Alina Wernick and Robert C. Williamson

478 Towards Justice-Oriented AI: Reflections on a Research-Community Partnership to Co-Develop Swahili AI Tools
Sybille Légitime, Kirabo Kakopo Atuhurira, Audrey Mugeni, Catherine D'Ignazio, Zawadi Amanya and Harini Suresh

497 What Drivers Want vs. What Drivers Take: A Worker-Centered View of Fairness in Ridehail Platforms
Samantha Dalal, Varun Nagaraj Rao, Christine Ao and Andrés Monroy-Hernández

510 When Interpretability Is Unequally Distributed: Fairness in Hybrid Interpretable Models
Ziba Jabbar Zare, Ulrich Aïvodji, Julien Ferry and Thibaut Vidal

525 The Benchmark Trap: Structures of Power and Injustice in AI Evaluations
Jason Branford and Angelie Kraft

545 Why and How People Check Generative AI Output for Mistakes
Patrick Gage Kelley, Derrick Feldmann, Reena Jana, Colleen Thompson-Kuhn and Allison Woodruff

551 Explainable AI for the EU Right to Explanation: A Systematic Review of the Law-XAI Translation Gap
Benjamin Fresz, Elena Dubovitskaya and Marco F. Huber

570 Bathtubs, Boundaries, and Sandboxes: AI Regulatory Learning Under Legal Uncertainty
Tom Deckenbrunnen, Alessio Buscemi, Marco Almada, Alfredo Capozucca and German Castignani

581 When LLM Governance Falls to Employees: Informal Practices of Responsible AI Use at Work
Tarini Saka, Sandy Hong and Veelasha Moonsamy

634 From Model Duties to Downstream Fundamental Rights: Contextual Accountability for GPAI Under the EU AI Act
Hannan Zhang and Xiangyi Liu

TUESDAY, OCTOBER 13

645 AI Adoption Across a Multinational Workforce: Sociotechnical Conditions for GenAI Acceptance in Human Resources
Dalia Ali, Maria José Rodríguez Velázquez, Manoel Horta Ribeiro, Q. Vera Liao and Orestis Papakyriakopoulos

661 AI-tribution: Enabling Training Data Attribution for Users of Generative Music Models
Julia Barnett, Nathan Pruyne, Nicholas Diakopoulos and Bryan Pardo

672 Prediction Laundering: The Sociotechnical Production of Truth on Polymarket
Yasaman Rohanifar, Sharifa Sultana and Syed Ishtiaque Ahmed

705 A Chain Is Only as Strong as Its Weakest Link: A Scoping Review of System Integration Audits in AI
Leah Davis, Dominic Martin and AJung Moon

725 AI Alignment Without Representation: From Scalar Rewards to Social Deliberation
Majid Ghasemi and Mark Crowley

748 Agentic Relationship Harm: Benchmarking and Gating Relational Manipulation in AI Agents
Tan Pei Sze, Tasuku Igarashi and Isao Echizen

768 Evaluating Occupational Bias in Anime-style Text-to-Image Generation: An MLLM-Based Workflow
Ayano Ito, Keiichi Namikoshi, Yuko Sakurai, Hiromi Arai and Satoshi Oyama

786 Beyond Alliance: Assessing the Utility of NATO's Principles for Responsible Use for AI in Defence for States Outside the Alliance
Daniel Trusilo

806 AI from the Margins (AIM): Rethinking Participatory AI Design Through the Lived Experience of Minoritized Communities
Tijs Portegies, Laureanne Willems, Maaike Harbers, Giovanni Sileno, Roland van Dierendonck, Mayesha Tasnim, Lotte Willemsen and Sennay Ghebreab

835 The Deliberative Deficit: An Empirical Critique of LLMs in Democratic Discourse
Maurice Flechtner

866 When AI Enters the Workplace, Who Faces Greater Risk? A Gendered Analysis
Miriam Fernandez, Ángel Pavón Pérez, Damiano Giallongo, Davide Ghia, Maryam Yaqub, Daniele Quercia and Tania Cerquitelli

897 Manipulation-Proof Oblivious Audits Against Deceptive Model Providers
Augustin Godinot, Sofiane Azogagh, Julien Ferry and Sébastien Gambs

910 Algorithmic Gender Prediction Is Illegitimate, but Gender Imputation Can Yield Valid Measurements
Evan Dong and Angelina Wang

919 Rethinking Robot Ownership: Comparing Public, Corporate, and Private Ownership Models
Isaac Sheidlower, Tomo Lazovich, Harini Suresh and Serena Booth

944 Who Brought Easter Eggs to Eid? Auditing LLM-Generated Cultural Translation of Math Word Problems Across Languages and Regions
Parisa Suchdev and Juniper Lovato

962 Lost in Translation: How Universal Ethical Values Fail to Translate Across Global Contexts
Ozioma C. Oguine, Munachimso B. Oguine, Cesar Cervera, Jenny Yang, Pooja Voladoddi, Mario Rodriguez, Saif Eddin Bani Malhem, Karla Badillo-Urquiola and Daricia Wilkinson

996 The Ownership Cost of AI Goal-Setting: Empirical Evidence for a Quality-Motivation Dissociation and Differential Harm to Vulnerable Users (Extended Abstract)
Vivienne Bihe Chi, Roman Rietsche, Andreas Göldi, Lyle Ungar and Sharath Chandra Guntuku

1025 Privacy Cost as Equity Input: A Group Fairness Criterion for Differentially Private Machine Learning
Rakshit Naidu

1051 Can LLMs Rank? A Tale of Triads and Triage
Gaurab Pokharel, Shafkat Farabi, Patrick J. Fowler and Sanmay Das

1071 The Disciplinary Language Transfer Problem: How Psychological Vocabulary Produces Governance Failures in AI Agent Deployment
Kymberly Lasser-Chere, Tyler Akidau and Marc Millstone

1121 RCTs & Human Uplift Studies: Methodological Challenges and Practical Solutions for Frontier AI Evaluation
Patricia Paskov, Kevin Wei, Shen Hong, Dan Bateyko, Xavier Roberts-Gaal, Carson Ezell, Gailius Praninskas, Valerie Chen, Umang Bhatt and Ella Guest

1122 MonitrLLM: A Community-Centered Evaluation Infrastructure for Large Language Models
Victor Ojewale, Ro Encarnación, Suresh Venkatasubramanian and Danaé Metaxa

1149 The Algorithmic Flattening of Sound: Computational Evidence and Justice Implications of AI Music Homogenization
Zoe Slendebroek and Danaé Metaxa

1202 Grounded but Misleading: Evaluating Semantic Alignment in AI-Generated Security Explanations
Heajun An, Connor Ng, Sandesh Sharma Dulal, Junghwan Kim and Jin-Hee Cho

1236 Characterizing Agentic Flooding of Government Services
Chris Schmitz, Lewis Hammond and Alan Chan

1308 Studying People to Study AI: Expert Perspectives on the Epistemic Fit and Barriers of Human Research in AI Safety & Ethics
Jessica Y. Bo, Paula Akemi Aoyagui, Shalaleh Rismani, Dipto Das, Syed Ishtiaque Ahmed and Ashton Anderson

1346 Fairness over Time: Evaluating How Fairness Objectives Shape Health Disparities in AI-Driven Clinical Decision Support Systems
Kaitlyn E. Wade, Alice C. W. Huang and Daniel J. Lizotte

1398 AI Literacy and Procedural Justice in Algorithmic Systems: A Conditional Account on Data Evaluation Practices and Human Oversight
Carmen Loefflad and Jens Grossklags

1438 Understanding Fairness–Privacy Interactions in Graph Neural Networks: A Literature Review, Taxonomy, and Open Challenges
Jingyan Zhang and Irina Tal

1450 Towards Characterising Conditions of Assistance, Hindrance and Depletion in Human Agency from AI Systems

WEDNESDAY, OCTOBER 14

8:30 AM **Keynote 3: Why can't we get it right? The challenges and limits of Responsible AI**

Irina Shklovski

Abstract: In 1987 Robert Kraut, then a computer scientist at Bell Labs, asked how can technology be designed "to exploit its usefulness without exploiting its users." Nearly four decades later, we still don't have an answer. Over the years arguments about technology have moved from concerns about data and privacy, to bias and discrimination in algorithmic systems, to ethical concerns about data driven AI systems. This culminated in a focus on "Responsible AI" – comprising of tools, methods, checklists, standards, compliance evaluations, and ways of thinking. Yet AI systems continue to fail us, plagued by the same problems of bias, privacy concerns, and overhyped promises that consistently fall short. Why do problematic AI systems seem unavoidable and what does it take to create AI "responsibly"? I will discuss the complex misalignment between conceptions of ethics and harm and the logics of modularity and scope that underpin technology development, the seemingly insurmountable yet persistently overlooked problems of data quality and the increasing challenge of moral stress in technical practice.

Bio: Irina Shklovski is a Professor of Communication and Computing in the Department of Computer Science at the University of Copenhagen. She also holds a WASP-HS visiting professorship in the Department of Thematic Studies, Gender Studies at Linköping University. Her research examines the relationship between technological infrastructures, institutional practices, and everyday human experience, focusing on responsible and ethical technology design, information privacy, creepy technologies and the sense of powerlessness people experience in the face of massive personal data collection. Current projects explore topics such as data quality, explainable AI, and moral stress in responsible AI development.

9:00 AM **Discussion Session 4: AI Across Contexts: Responsibility, Values, and Power**
Chair: Zee Talat

548 AI Alignment and Fiduciary Obligation
Benjamin Lange

1173 Abstracted Away: Resisting Alienation and Ungrounded Abstraction in AI Research Communities
Vyoma Raman, Isabel Gallegos and Neha Srivathsa

1246 The Judgment-Consequence Gap: LLM Moral Reasoning in Healthcare Decisions
Hadi Hosseini, Samarth Khanna and Leona Pierce

1374 Culturally Situated AI Safety for Youth: Saudi Arabian Perspectives of Youth, Parents and Teachers
Aljawharah Alzahrani and Tanusree Sharma

10:15 AM **Poster Session 3 and Refreshment Break**

WEDNESDAY, OCTOBER 14

10:45 AM **Paper Session 5: Auditing and Evaluating AI**

Chair: Mark Diaz

471 No More Plaudits for Algorithm Audits: A Provocation
Alicia DeVrio and David Widder

751 Frictionless Validation: Exploring AI Sycophancy and Developmental Risks
Junfeng Jiao, Abhejay Murali and Saleh Afroogh

1010 The Effects of Functional Emotions on LLM-Based Paper Review
Pedro Bento, Yan Aquino Amorim, Arthur Rodrigues Chagas, Arthur Buzelin, Lucas Dayrell, Gisele Lobo Pappa, Virgilio Almeida and Wagner Meira Jr.

1090 Auditing AI-Washing in German Startups: A Mixed-Methods Study of Marketing Claims and Perceptions
A Pranav, Luke Hartmann and Anne Lauscher

1152 What Current AI Benchmarks Leave Unmeasured: Modality, Search, Citations, and Implications (for Safety Evaluations)
Ro Encarnación, Tina Behzad, Emma Lurie and Danaé Metaxa

12:00 PM **Informal Town Hall**

12:15 PM **Lunch Break (on your own)**

1:30 PM **Keynote 4: Thoughts for the AI and AI Ethics Community, from a Computer Security Research Perspective**

Tadayoshi Kohno

Abstract: The computer security research field: (1) identifies how computer systems can be used to harm others and (2) develops methods to prevent or mitigate those harms. In the computer security field, the harms are caused by adversaries. In the AI field, in some cases, the harms are caused by the technology's mere existence. Still, there are parallels. In this talk, I reflect on a career of computer security research and education and offer thoughts and lessons for the AI and AI ethics community as it, too, wrestles with identifying and mitigating harms.

Bio: Tadayoshi Kohno (Yoshi) is the McDevitt Chair in Computer Science, Ethics, and Society and professor in the Department of Computer Science and the Center for Digital Ethics at Georgetown University. His research focuses on helping protect the security, privacy, and safety of users of current and future technologies. He is a recipient of the Alfred P. Sloan Research Fellowship, the U.S. National Science Foundation CAREER Award, the Technology Review TR-35 Young Innovator Award, and the Golden Goose Award. He has received five Test of Time Awards for his research. Kohno serves as a member of the Electronic Frontier Foundation Board of Directors, the Max Planck Institute for Security and Privacy Scientific Advisory Board, and the USENIX Security Steering Committee. He was an inaugural member of the U.S. National Academies Forum on Cyber Resilience. Kohno was previously a professor in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, where he co-directed the University of Washington Computer Security & Privacy Research Lab and was the Associate Dean for Faculty Success for the University of Washington College of Engineering. He is a Fellow of the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronics Engineers (IEEE).

WEDNESDAY, OCTOBER 14

2:00 PM **Paper Session 6: Generative AI and Human Vulnerability**
Chair: Daniel Susser

623 Engagement-Optimized Care: When LLMs Become Mental Health Infrastructure
Briana Vecchione, Meryl Ye, Livia Garofalo and Ranjit Singh

908 Beyond Guardrails: Dual Failure Modes in LLM Safety as Inverse Control-System Errors
Ethan Gillese

1057 Personality Without Persons? A Psychometric Critique of Big Five Testing in Large Language Models
Kim Zierahn, Cristina Cachero, Anna Korhonen and Nuria Oliver

1079 The Beginning of ChatGPT Ads
Emma Lurie, Ro Encarnación, Sorelle A. Friedler and Danaé Metaxa

1107 AI-Enabled Human Memory Manipulation: Misleading AI-Generated Summaries Distort Human Memory
Mattea Sim, Yael Eiger and Tadayoshi Kohno

3:15 PM **Poster Session 3 (cont.) and Refreshment Break**

Posters

79 Enabling the 1:1 Promise: AI Enabled Fairness for Indonesia's Retail IPO Allocation
Anchal Gupta, Gleb Papyshev and James T Kwok

123 'We're Not Doing Ethical AI, We're Doing Legal AI': How Compliance Shapes Responsible AI Practice
Tallal Ahmad, Arshia Dutta, Corey Jackson and Abraham Mhaidli

124 Epistemology Gives a Future to Complementarity in Human-AI Interactions
Andrea Ferrario, Alessandro Facchini and Juan M. Duran

199 Is This AI? Longitudinal Analysis of Strategies Used for AI Detection on Two Subreddits
Christina Yeung, Galen Weld, Jaron Mink and Franziska Roesner

201 Examining Risks Through a Characterization of the AI Companion Application Ecosystem: A Stratified Sample from the Apple App Store and Google Play Store
Natalie Grace Brigham, Lucy Qin and Tadayoshi Kohno

204 Can LLM Code Explanations Adapt to Diverse Problem-Solvers' Needs?
Andrew Anderson, David Piorkowski, Justin D. Weisz, Margaret Burnett and Kush R. Varshney

249 On Seeding Watermarks to Detect Verbatim LLM Copy-Paste Responses
Aizierjiang Aiersilan, Artin Yousefi and Robert Pless

293 Tracking AI Advisory Drift Before It Reaches the Farms: Measuring Contextual Reliability in LLM-Generated Advice Across the Agricultural Landscape
Jerry H. Martin and Tianyi Li

306 Beyond Robot Ownership: A Lease-Based Framework for Lifecycle Governance of High-Risk Embodied AI (Extended Abstract)
Yiqing Sun

WEDNESDAY, OCTOBER 14

314 Safety Drift After Fine-Tuning: Evidence from High-Stakes Domains
Emaan Bilal Khan, Amy Winecoff, Miranda Bogen and Dylan Hadfield-Menell

353 Parental Notification Gaps in ChatGPT-5: An Empirical Study
Kerem Ersoz, Saleh Afroogh, David Atkinson and Junfeng Jiao

360 LLMs as Narrators of Lived Experience: A Study of Generative Hermeneutical Epistemic Injustice in ADHD Sense-Making
Anne Arzberger and Dongying Zhang

400 Synthetic Sources?: Auditing Generative Search Engine Citations for Evidence of AI-Generated Sources
Mowafak Allaham and Nicholas Diakopoulos

432 Secure Human Oversight of AI: Threat Modeling in a Socio-Technical Context
Jonas C. Ditz, Veronika Lazar, Elmar Lichtmeß, Carola Plesch, Matthias Heck, Kevin Baum and Markus Langer

447 Identifying Harm in Personalized, Generative AI Systems Requires User-Centered Auditing at the Interaction Level
Hannah Cha

483 Selective Overconfidence: How Large Language Models Underserve Social Welfare Law Across Jurisdictions
Shera Potka, Alex Thomo, Paula Helm and Wulf Loh

488 FedFAn: Federated Learning for Facial Animation
Jack Parry and Vinay Namboodiri

501 The Limits of Machine Unlearning as a Means of Political Protest
Houssam Abbas and Alicia Patterson

516 Test-Time Collective Action: Proxy-Based Perturbations for Correcting Algorithmic Harms
Meghana Bhange, Ulrich Aïvodji and Elliot Creager

527 Five Obstacles to AI Trustworthiness Alignment and How to Deal with Them: Introducing the DELIB-TRUST Framework
Pascal Koenig and Siddharth Mehrotra

529 Rules or Character? Scaling Laws for AI Safety Design (Extended Abstract)
Satoshi Takahashi, Nobuji Kouno, Masaaki Komatsu and Ryuji Hamamoto

547 Are Algorithm Registers Transparent? Perspectives from Germany
Iman Peljto, Xenia Heilmann and Mattia Cerrato

562 Missing Melodies: AI Music Generation and Its "Nearly" Complete Omission of the Global South
Atharva Mehta, Shivam Chauhan and Monojit Choudhury

577 Towards Fair and Accurate Short VAAs: Adaptive Ordering and Imputation in Democratic Recommendation Systems
Thilo I. Dieing and Lea Cohausz

614 A Critical Pragmatism Approach for AI Ethics: Lessons from Urban Planning Theory
Jennah Gosciak, Karen Levy and Allison Koenecke

WEDNESDAY, OCTOBER 14

624 AI Revealed Preferences
Peter Salib, Sam Wang, Sofiia Lobanova, Yonathan Arbel and Simon Goldstein

627 Improving AI Guard Rails with Moral Trust
Taylor Olson

635 Examining the Effect of Explanations of AI Privacy Redaction in AI-mediated Interactions
Roshni Kaushik, Maarten Sap and Koichi Onoue

640 Protecting the Underrepresented: Data Debiasing via Asymmetric Adversarial Encoding for Fair Machine Learning
Deepanshu Aggarwal and Puspita Majumdar

647 Auditing LLM Portrayals of Neurodivergent People: Quantifying the Asymmetry Between Deficit Framing and Neurodiversity Affirmation
Carlos Toxtli and Manuel Delaflor

662 From Forensics to Ecosystems: Rethinking Watermarks for Generative AI Oversight
Daniel Susser, John Thickstun and Gili Vidan

665 Check and Trust with PAU: A Paper Analysis Utility for Automated Bibliographic Verification
Pauline Speckmann, Raphael Fischer and Christian Janiesch

681 AI-Enabled Authoritarian Practices in Formal Democracies
Keir Reid, Isaac Friend and Fazl Barez

711 When Do AI Gains Become Broadly Shareable? A Policy Threshold for AI-Driven Automation
Aran Nayebi

726 Fairness-Aware Optimization in Logistic Regression: Empirical Trade-Offs Across Healthcare and Socioeconomic Domains
David Oluwabukunmi Shodipo and Blessing Ogbuokiri

761 Towards Stakeholder-Grounded Requirements Engineering in AI System Specification
Danniell Hu, Westley Weimer and Elizabeth Bondi-Kelly

771 SurakshaEval: An Indic Safety Benchmark for Multilingual LLMs
Debopriyo Banerjee, Kapil Rajesh Kavitha, Angana Borah, Xudong Han, Yuxia Wang, Parameswari Krishnamurthy, Utkarsh Agarwal, Atharva Kulkarni, Swaran Lata, Ayush Munot, Dhruv Sahnan, Aaryamonvikram Singh, Preslav Nakov and Monojit Choudhury

782 People Are Not Just Their Countries. Disentangling Social Determinants of LLM Value Alignment Across Europe
Maria-Louisa Wightman, Guillaume Bied and Tijl De Bie

791 Personalizing AI to Align with the Better You
Jakob Stenseke, Jocelyn Zhu and Dylan Hadfield-Menell

815 Curating the Collapse: How Sample-Level Auditing Accelerates Model Collapse in Synthesis of Tabular Data
William Martin Stevenson, Anne Gerdes, Lukas Galke Poech and Anton Danholt Lautrup

831 Hardware Is an AI Ethics Problem: Expert Visions for a Sustainable and Equitable Semiconductor Industry
Naira Paola Arnez Jordan, Chiara Ullstein, Michel Hohendanner, Jens Grossklags, Lorenzo Servadei, Alejandro Merino-Madrid and Orestis Papakyriakopoulos

WEDNESDAY, OCTOBER 14

841 Care and Capability: How to Temper Excessive Social-Emotional Chatbot Use According to Analogous Professional Practice
Tony Wang and Qian Yang

883 Quantifying System-Level Harms from AI Adoption in Complex Sociotechnical Systems
Paul Vautravers, Oliver Chalkley, Gabriel Downer, Kate S and Damian Ruck

898 Shaping the Future of Generative AI for Black Communities: A Frame Analysis of Public Discourse and Empirical Scholarly Research
Angela D. R. Smith, Gabriella Thompson, Christopher L. Dancy, Mark Diaz, Seyi Olojo and Christina N. Harrington

916 Double Disadvantage: An Interview Study with Women in the AI Industry on the Role of Gender in AI Ethics
Maaike Harbers, Kirsten van der Stelt and Sander 't Sas

959 The AI Assessment Sandbox Configurator: A Framework to Support Technical Assessment in AI Regulatory Sandboxes
Alessio Buscemi, German Castignani, Daniele Pagani, Maxime Cordy and Jordi Cabot

980 The Hidden Tax of AI Safety: Residual Risk Redistribution in Automated Governance
Henrietta Boowuo and Ernest Ofori-Mensah

983 Deconstructing AI Hype: A Playbook for Resistance
Shreya Chowdhary, Justine Zhang and Ben Green

1009 Let's Unlearn Stereotypes Before Decision-Making: The Impact of Intrinsic Bias Mitigation on Downstream Fairness in LLMs
Mina Arzaghi, Alireza Dehghanpour Farashah, Florian Carichon, Jean-François Plante and Golnoosh Farnadi

1030 Knowable Errors: Architectural Foreseeability and the Case for AI Recklessness in Safety-Critical Deployment
Fawaz Ishola

1054 Data Annotation as Measurement
Emma Harvey, Allison Koenecke and Rene F. Kizilcec

1058 Simulation as Alibi—How the Social Order Gets Outsourced to LLM Agents
Tang Shiyun and Lin Zezheng

1073 A Taxonomy of AI Value Alignment: Subproblems and Solutions
Daniel R. Harris, Karina Vold, Harsh Kumar and Ashton Anderson

1089 Critical Acclaim Orientation in Large Language Models: Evidence from Film Preference Elicitation
Jonghyun Jee and Aaron Shaw

1123 Designing for Doubt: The Case for Informed Abstention in Autonomous Agents
Victor Ojewale and Suresh Venkatasubramanian

1171 Agentic AI: User Empowerment or Foreclosure?
David Gamba, Daniel M. Romero and Grant Schoenebeck

WEDNESDAY, OCTOBER 14

1195 Open Science, Closed Models: How Funding Shapes AI in Science
Ana Trišovi· and Janakan Sivaloganathan

1213 Who Bears Privacy Risk? Model-Dependent Memorization in Synthetic Tabular Data
Aidan LoStracco, Joshua August Skorburg and Wenjing Zhang

1232 From a River in Gilead to the Inference Distributions of Large Language Models: Covert Dialect Bias and Linguistic Profiling at Scale
Chowdhury Mohammad Abdullah and Rita Orji

1261 Between Commons and Property: Individual Image-Generation Model Creators' Perceived Accountability for Derivative AI Model on Open Model Marketplaces
Eun Jeong Kang, Fengyang Lin and Angel Hsing-Chi Hwang

1262 Silent Updates: Measuring and Closing the Post-Deployment Disclosure Gap
Sophia Abraham and Ben Bucknall

1303 Model Multiplicity and Predictive Arbitrariness in Recidivism Risk Assessment
Ashwin Singh and Carlos Castillo

1314 Epistemic Framing Tilt in LLMs: Auditing Religious Discourse Across Traditions
Arthur Buzelin, Yan Aquino, Pedro Bento, Arthur Chagas, Lucas Dayrell, Gisele L. Pappa, Virgilio Almeida and Wagner Meira Jr.

1383 The Evaluation Differential: When Frontier AI Models Recognise They Are Being Tested
Varad Vishwarupe, Nigel Shadbolt and Marina Jirotka

1394 When Government Uses AI: Experimental Evidence of Selective Trust Decline and Algorithmic Nimbyism
Daniel Juhász Vigild, Andreas Bjerre-Nielsen, Laust Hvas Mortensen and Vedran Sekara

1404 Negotiating Risk Boundaries in AI for Policing Through Mixed-Stakeholder Deliberation
Mackenzie Jorgensen, Jo Reilly, Alex Sutherland and Miri Zilka

1408 Emergent Languages in Populations of Language Model Agents: From Token Efficiency to Oversight Evasion
Stine Lyngsø Beltoft, William Brach, Federico Torrielli, Jacob Nielsen, Annemette Brok Pirchert, Filippo Tonini, Peter Schneider-Kamp and Lukas Galke Poech

1439 Do Compliance Judges Read the Rules? READ: Robustness Through Erasure, Ablation and Diagnosis of LLM Policy Grounding
Adam Kahirov, Bastian Lipka, Jasmina Gajcin and Elizabeth Daly

1449 Gender Bias and Sonic Gender Performance in AI-Generated Music: Representation, Expression, and Perception
Minyeong Bae, Hyunjae Kim and Kyung Myun Lee

1462 The Geography of Algorithmic Judgment: LLM Intermediaries, Place Identity, and Racial Steering in Housing Search
Hana Samad, Trung Lam, Christoph Mügge-Durum and Michael Akinwumi

4:00 PM **Closing Remarks and Awards**

AIES-26 SPONSORS

AIES 2026 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.

GOLD

Google

SILVER

IBM

BRONZE

AI Policy Lab
at Umeå University

ERICSSON

150 years of reaching new horizons

TECH POLICY LAB
UNIVERSITY of WASHINGTON

ORGANIZED BY

Association for the Advancement of Artificial Intelligence

acm Association for Computing Machinery