THE PROGRAM GUIDE FOR

Sixth AAAI/ACM Conference on

ARTIFICIAL INTELLIGENCE, ETHICS, & SOCIETY

AUGUST 8–10, 2023 | MONTRÉAL, CANADA

Follow AIES 2023 on Twitter! @AIESConf

AIES-23 WI-FI ACCESS

Network Name

Palais_des_congres_de_Montreal

Password

PCM34602

TRAVELING TO MONTRÉAL

Tourisme Montréal

PROGRAM CONTENTS

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

CONFERENCE CHAIR

Francesca Rossi *IBM*

CONFERENCE PROGRAM CO-CHAIRS

Sanmay Das *George Mason University*
Kay Firth-Butterfield *World Economic Forum*
Alex John London *Carnegie Mellon University*
Jenny Davis *Australian National University*

LOCAL CHAIR

Marc-Antoine Dilhac *Mila*

STUDENT PROGRAM CHAIRS

Brent Venable *Tulane University*
Su Lin Blodgett *Microsoft Research*
Theodore Lechterman *IE University*
Wenbin Zhang *Michigan Technological University*

WORKFLOW CHAIRS

Tasfia Mashiat *George Mason University*
Gaurab Pokharel *George Mason University*

The AIES-23 Program Committee welcomes you to Montréal!

WELCOME FROM THE
CONFERENCE CHAIR

AI is increasingly pervasive and powerful, and has the potential to empower individuals and improve society. However, the ethical ramifications of AI systems and their impact on human societies requires a deep multi-disciplinary socio-technical reflection. The AIES conference, now at its 6th edition, has pursued this mission at the scientific level, engaging both senior and junior researchers and covering all dimensions of AI ethics, including technical advances, policies, philosophy, and economics. I am very excited about this year's program, and I am looking forward to meeting all the attendees in Montréal. I would also like to deeply thank all our sponsors for their generous support to AIES 2023.

Francesca Rossi *IBM*

WELCOME FROM THE
PROGRAM CO-CHAIRS

The AIES conference is convened each year by the AIES steering committee and its technical program is designed by our program co-chairs from Computer Science, Law and Policy, the Social Sciences, Ethics and Philosophy. Our goal is to encourage talented scholars in these and related fields to present and discuss the best work related to morality, law, policy, psychology, the other social sciences, and AI. In addition to the community of scholars who have participated in these discussions from the outset, we explicitly welcome disciplinary experts who are newer to this topic, and see ways to break new ground in their own fields by thinking about AI. In this context, we are very excited about this year's program and welcome you to AIES 2023!

Organizing AIES would not have been possible without the contributions of many people. Francesca Rossi has been a model leader as Conference Chair. Theodore Lechterman, Su Lin Blodgett, Wenbin Zhang, and Brent Venable have put tremendous energy into organizing the student program. Gaurab Pokharel and Tasfia Mashiat were of great help in the innumerable tasks involved in organizing the paper reviewing and selection process. Francisco Cruz provided invaluable support for our web presence. Marc-Antoine Dilhac has graciously helped us with local organization and involving

Mila, Vince Conitzer was always available for advice, and none of this would have been possible without the incredible organizational abilities of Meredith Ellison and Chesley Grove at AAAI.

Finally, the biggest thanks must go to the authors who submitted papers, the program committee members who spent countless hours thoughtfully reviewing them, as well as the broader AIES community, who keep working on and thinking about the important questions. We are grateful for this opportunity to support the community in its goals, and look forward to sharing an experience in Montreal that is both intellectually rich and of genuine importance in the world today.

Sanmay Das *George Mason University* · **Kay Firth-Butterfield** *World Economic Forum*
Alex John London *Carnegie Mellon University* · **Jenny Davis** *Australian National University*

WELCOME FROM THE
STUDENT PROGRAM CHAIRS

Addressing the ethical challenges and opportunities of AI will occupy many generations to come. AIES is especially delighted to be able to support the next generation of AI ethics researchers. Our student track provides promising Ph.D. students with targeted programming, mentorship, and financial support to attend and share their work at AIES. We are pleased this year to welcome students from 5 continents working across STEM, social science, and the humanities. We thank the conference and program chairs for making the student track a central component of AIES and are grateful to our funders for helping to make the AIES experience accessible to junior researchers.

Brent Venable *Tulane University* · **Su Lin Blodgett** *Microsoft Research*
Theodore Lechterman *IE University* · **Wenbin Zhang** *Michigan Technological University*

PROGRAM OVERVIEW

TUESDAY, AUGUST 8

9:00–9:10
Welcome

9:10–10:15
Opening Keynote

10:15–10:35
Break

10:35–12:30
Proceedings Session 1

12:30–2:00
Lunch
(Student Track convenes separately)

2:00–3:00
Proceedings Session 2

3:00–4:10
Keynote Panel

4:10–4:30
Break

4:30–6:00
Proceedings Session 3

6:00–8:00
Poster Session 1 + Reception

WEDNESDAY, AUGUST 9

9:00–10:00
Panel

10:00–10:15
Awards Announcements

10:15–10:35
Break

10:35–12:10
Proceedings Session 4

12:10–2:00
Lunch

2:00–3:00
Proceedings Session 5

3:00–4:10
Keynote 2

4:10–4:30
Break

4:30–6:00
Proceedings Session 6

6:00–8:00
Poster Session 2 + Reception

THURSDAY, AUGUST 10

9:00–9:40
Proceedings Session 7

9:40–11:40
Poster Session 3

11:40–12:50
Closing Keynote

12:50–1:00
Conference Closing

TUESDAY, AUGUST 8

9:00–9:10 **Welcome Remarks // Room 510AC**
Francesca Rossi, IBM
Sanmay Das, George Mason University
Theodore Lechterman, IE University

9:10–10:15 **Opening Keynote // Room 510AC**

The Generative AI Deployment Rush: How to Democratize the Politics of Pace
Annette Zimmerman

10:15–10:35 **Short Break**

10:35–12:30 **Proceedings Session 1 // Room 510AC**
Session Chair: Aylin Caliskan

Talks

Protecting Children from Online Exploitation: Can a trained model detect harmful communication strategies?
Darren Cook, Miri Zilka, Heidi Desandre, Susan Giles, and Simon Maskell

Analysis of Climate Campaigns on Social Media using Bayesian Model Averaging
Tunazzina Islam, Ruqi Zhang, and Dan Goldwasser

AI Art and Misinformation* Approaches and strategies for media literacy and fact checking
Johanna Walker, Gefion Thuermer, Elena Simperl, and Julian Vincens

From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research
Michael Feffer, Michael Skirpan, Hoda Heidari, and Zachary Lipton

How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making?
Saumik Narayanan, Guanghui Yu, Chien-Ju Ho, and Ming Yin

Student Program Lightning Talks

Exploring the Effect of AI Assistance on Human Ethical Decisions
Saumik Narayanan

Queering Futures with Data-Driven Speculation the design of an expanded-mixed methods framework integrating qualitative, quantitative, and practice-based modes
Jess Westbrook

Are Model Explanations Useful in Practice? Rethinking How to Support Human-ML Interactions
Valerie Chen

Ranked Candidate Fairness in Preference Aggregation
Kathleen Cachel

The ELIZA Defect: Constructing the Right Users for Generative AI
Daniel Affsprung

Governing Silicon Valley and Shenzhen: Assessing a New Era of Artificial Intelligence Governance in the US and China
Emmie Hine

Lightning Talks

User Tampering in Reinforcement Learning Recommender Systems
Charles Evans and Atoosa Kasirzadeh

Beyond the ML Model: Applying Safety Engineering Frameworks to Text-to-Image Development
Shalaleh Rismani, Renee Shelby, Andrew Smart, Renelito Delos Santos, Ajung Moon, and Negar Rostamzadeh

Reward Reports for Reinforcement Learning
Thomas Gilbert, Nathan Lambert, Sarah Dean, Tom Zick, and Aaron Snoswell

A Systematic Review of Ethical Concerns with Voice Assistants
William Seymour, Xiao Zhan, Mark Coté, and Jose Such

The Ethical Implications of Generative Audio Models: A Systematic Literature Review
Julia Barnett

Robust Artificial Moral Agents and Metanormativity
Tyler Cook

Mitigating Voter Attribute Bias for Fair Opinion Aggregation
Ryosuke Ueda, Koh Takeuchi, and Hisashi Kashima

Learning Optimal Fair Decision Trees: Trade-offs Between Interpretability, Fairness, and Accuracy
Nathanael Jo, Sina Aghaei, Jack Benson, Andres Gomez, and Phebe Vayanos

Model Debiasing via Gradient-based Explanation on Representation
Jindi Zhang, Luning Wang, Dan Su, Yongxiang Huang, Caleb Chen Cao, and Lei Chen

Sampling Individually-Fair Rankings that are Always Group Fair
Sruthi Gorantla, Anay Mehrotra, Amit Deshpande, and Anand Louis

Keep Sensors in Check: Disentangling Country-Level Generalization Issues in Mobile Sensor-Based Models with Diversity Scores
Alexandre Nanchen, Lakmal Meegahapola, William Droz and Daniel Gatica-Perez

A Deep Dive into Dataset Imbalance and Bias in Face Identification
Valeriia Cherepanova, Steven Reich, Samuel Dooley, Hossein Souri, John Dickerson, Micah Goldblum, and Tom Goldstein

TUESDAY, AUGUST 8

12:30–2:00 **Lunch Break (on own)**

2:00–3:00 **Proceedings Session 2 // Room 510AC**
Session Chair: Alexandra Olteanu

Talks

Iterative Partial Fulfillment of Counterfactual Explanations: Benefits and Risks
Yilun Zhou

Multicalibrated Regression for Downstream Fairness
Ira Globus-Harris, Varun Gupta, Christopher Jung, Michael Kearns, Jamie Morgenstern, and Aaron Roth

Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models
Peter Henderson, Eric Mitchell, Christopher Manning, Dan Jurafsky, and Chelsea Finn

Not So Fair: The Impact of Presumably Fair Machine Learning Models
Mackenzie Jorgensen, Hannah Richert, Elizabeth Black, Natalia Criado, and Jose Such

3:00–4:00 **Keynote Panel // Room 510AC**
Moderator: Alex John London

Large Language Models: Hype, Hope, and Harm
Roxana Daneshjou, Atoosa Kasirzadeh, Kate Larson, and Gary Marchant

4:10–4:30 **Short Break**

4:30–6:00 **Proceedings Session 3 // Room 510AC**
Session Chair: Patrick Fowler

Talks

Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare
Eran Tal

Unpicking Epistemic Injustices in Digital Health: On the Implications of Designing Data-Driven Technologies for the Management of Long-Term Conditions
S J Bennett, Caroline Claisse, Ewa Luger, and Abigail C. Durrant

Evaluating the Impact of Social Determinants on Health Prediction in the Intensive Care Unit
Ming Ying Yang, Gloria Hyunjung Kwak, Tom Pollard, Leo Anthony Celi, and Marzyeh Ghassemi

Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI
Hubert D. Zajac, Natalia R. Avlona, Finn Kensing, Tariq O. Andersen, and Irina Shklovski

Student Program Lightning Talks

The Role of Governance in Bridging AI Responsibility Gaps
Bhargavi Ganesh

Algorithm-Assisted Decision Making and Racial Disparities in Housing: A Study of the Allegheny Homelessness Assessment Tool
Lingwei Cheng

Designing Interfaces to Elicit Data Issues for Data Workers
Kevin Bryson

True and Fair: Robust and Unbiased Fake News Detection via Interpretable Machine Learning
Chahat Raj

Benchmarked Ethics: A Roadmap to AI Alignment, Moral Knowledge, and Control
Aidan Kierans

Lightning Talks

AI Art and its Impact on Artists
Harry Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru

Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms
Nathan Dennler, Anaelia Ovalle, Ashwin Singh, Luca Soldaini, Arjun Subramonian, Huy Tu, William Agnew, Avijit Ghosh, Kyra Yee, Irene Font Peradejordi, Zeerak Talat, Mayra Russo, and Jess De Jesus De Pinho Pinhal

Action Guidance and AI Alignment
Pamela Robinson

Ethical and Social Risks of Generative Text-to-Image Models
Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh

On the Connection between Game-Theoretic Feature Attributions and Counterfactual Explanations
Emanuele Albini, Shubham Sharma, Saumitra Mishra, Danial Dervovic, and Daniele Magazzeni

Adaptive Adversarial Training Does Not Increase Recourse Costs
Ian Hardy, Jayanth Yetukuri, and Yang Liu

REFRESH: Responsible and Efficient Feature Reselection guided by SHAP values
Shubham Sharma, Sanghamitra Dutta, Emanuele Albini, Freddy Lecue, Daniele Magazzeni, and Manuela Veloso

Fairness Implications of Encoding Protected Categorical Attributes
Carlos Mougan, Jose M. Alvarez, Salvatore Ruggieri, and Steffen Staab

Diffusing the Creator: Attributing Credit for Generative AI Outputs
Donal Khosrowi, Finola Finn, and Elinor Clark

TUESDAY, AUGUST 8

6:00–8:00 **Poster Session & Reception // Room 510BD**

Posters

User Tampering in Reinforcement Learning Recommender System
Charles Evans and Atoosa Kasirzadeh

Beyond the ML Model: Applying Safety Engineering Frameworks to Text-to-Image Development
Shalaleh Rismani, Renee Shelby, Andrew Smart, Renelito Delos Santos, Ajung Moon, and Negar Rostamzadeh

Reward Reports for Reinforcement Learning
Thomas Gilbert, Nathan Lambert, Sarah Dean, Tom Zick, and Aaron Snoswell

A Systematic Review of Ethical Concerns with Voice Assistants
William Seymour, Xiao Zhan, Mark Coté, and Jose Such

The Ethical Implications of Generative Audio Models: A Systematic Literature Review
Julia Barnett

Robust Artificial Moral Agents and Metanormativity
Tyler Cook

Mitigating Voter Attribute Bias for Fair Opinion Aggregation
Ryosuke Ueda, Koh Takeuchi, and Hisashi Kashima

Learning Optimal Fair Decision Trees: Trade-offs Between Interpretability, Fairness, and Accuracy
Nathanael Jo, Sina Aghaei, Jack Benson, Andres Gomez, and Phebe Vayanos

Model Debiasing via Gradient-based Explanation on Representation
Jindi Zhang, Luning Wang, Dan Su, Yongxiang Huang, Caleb Chen Cao, and Lei Chen

Sampling Individually-Fair Rankings that are Always Group Fair
Sruthi Gorantla, Anay Mehrotra, Amit Deshpande, and Anand Louis

Keep Sensors in Check: Disentangling Country-Level Generalization Issues in Mobile Sensor-Based Models with Diversity Scores
Alexandre Nanchen, Lakmal Meegahapola, William Droz, and Daniel Gatica-Perez

A Deep Dive into Dataset Imbalance and Bias in Face Identification
Valeriia Cherepanova, Steven Reich, Samuel Dooley, Hossein Souri, John Dickerson, Micah Goldblum, and Tom Goldstein

AI Art and its Impact on Artists
Harry Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru

Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms
Nathan Dennler, Anaelia Ovalle, Ashwin Singh, Luca Soldaini, Arjun Subramonian, Huy Tu, William Agnew, Avijit Ghosh, Kyra Yee, Irene Font Peradejordi, Zeerak Talat, Mayra Russo, and Jess De Jesus De Pinho Pinhal

Diffusing the Creator: Attributing Credit for Generative AI Outputs
Donal Khosrowi, Finola Finn, and Elinor Clark

Student Posters

Exploring the Effect of AI Assistance on Human Ethical Decisions
Saumik Narayanan

Queering Futures with Data-Driven Speculation the design of an expanded-mixed methods framework integrating qualitative, quantitative, and practice-based modes
Jess Westbrook

Are Model Explanations Useful in Practice? Rethinking How to Support Human-ML Interactions
Valerie Chen

Ranked Candidate Fairness in Preference Aggregation
Kathleen Cachel

The ELIZA Defect: Constructing the Right Users for Generative AI
Daniel Affsprung

Governing Silicon Valley and Shenzhen: Assessing a New Era of Artificial Intelligence Governance in the US and China
Emmie Hine

The Role of Governance in Bridging AI Responsibility Gaps
Bhargavi Ganesh

Algorithm-Assisted Decision Making and Racial Disparities in Housing: A Study of the Allegheny Homelessness Assessment Tool
Lingwei Cheng

Designing Interfaces to Elicit Data Issues for Data Workers
Kevin Bryson

WEDNESDAY, AUGUST 9

9:00–10:00 **Panel Discussion // Room 510AC**

Moderator: Sanmay Das

AI *for* Society: Developing, Deploying, and Auditing Public-Facing AI
Patrick J. Fowler, Hoda Heidari, and Christo Wilson

10:00–10:15 **Awards Announcements // Room 510AC**

10:15–10:35 **Short Break**

10:35–12:10 **Proceedings Session 4 // Room 510AC**

Session Chair: Ajung Moon

Talks

Machine Learning practices and infrastructures
Glen Berman

Fairness Toolkits, A Checkbox Culture? On the Factors that Fragment Developer Practices in Handling Algorithmic Harms
Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju

Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness
Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang

A multidomain relational framework to guide institutional AI research and adoption
Vincent Straub, Deborah Morgan, Youmna Hashem, John Francis, Saba Esnaashari, and Jonathan Bright

Student Program Lightning Talks

Towards a Holistic Approach: Understanding Sociodemographic Biases in NLP Models using an Interdisciplinary Lens
Pranav Venkit

Algorithmic Bias: When stigmatization becomes a perception
Olalekan Akintande

Ethical Principles for Reasoning about Value Preferences
Jessica Woodgate

Explainability in Process Mining: A Framework for Improved Decision-Making
Luca Nannini

Can AlphaGo be apt subjects for Praise/Blame for “Move 37”?
Mubarak Hussain

Anticipatory regulatory instruments for AI systems — A comparative study of regulatory sandbox schemes
Deborah Morgan

Examining the Ethics of Brain-Computer Interfaces: Ensuring Safety, the Rights and Dignity of Personhood
Terkura Thomas Mchia

Lightning Talks

Flickr Africa: Examining Geo-Diversity in Large-Scale, Human-Centric Visual Data
Keziah Naggita, Julienne LaChance, and Alice Xiang

Evaluation of targeted dataset collection on racial equity in face recognition
Rachel Hong, Tadayoshi Kohno, and Jamie Morgenstern

Evaluating Biased Attitude Associations to Language Models in an Intersectional Context
Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan

Unmasking Nationality Bias: A Study of Human Perception of Nationalities in AI-Generated Articles
Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson

No Justice, No Robots: From the Dispositions of Policing to an Abolitionist Robotics
Tom Williams and Kerstin Sophie Haring

Why We Need to Know More: Exploring the State of AI Incident Documentation Practices
Violet Turri and Rachel Dzombak

What does it mean to be a responsible AI practitioner: An ontology of roles and skills
Shalaleh Rismani and Ajung Moon

12:10–2:00 **Lunch Break (on own)**

2:00–3:00 **Proceedings Session 5 // Room 510AC**
Session Chair: Jesse Kirkpatrick

Talks

Effective Enforceability of EU Competition Law Under AI Development Scenarios: a Framework for Anticipatory Governance
Shin-Shin Hua and Haydn Belfield

The Bureaucratic Challenge to AI Governance: An Empirical Assessment of Implementation at U.S. Federal Agencies
Christie Lawrence, Isaac Cui, and Daniel Ho

Self-determination through explanation: an ethical perspective on the implementation of the transparency requirements for recommender systems set by the Digital Services Act of the European Union
Matteo Fabbri

WEDNESDAY, AUGUST 9

Student Program Lightning Talks

Public attitudes toward ethical AI in courts: A Vignette Survey and Deliberation Experiment
Arna Woemmel

Sealed Knowledges: A Critical Approach to the Usage of LLMs as Search Engines
Nora Freya Lindemann

Investigating the Relative Strengths of Humans and Machine Learning in Decision-Making
Charvi Rastogi

A Responsible Artificial Intelligence Approach to Epistemic Credibility in Artificial Intelligence Systems
Abiola Azeez

Safety in Conversational Systems
Jinhwa Kim

Lightning Talks

Reckoning with the Disagreement Problem: Explanation Consensus as a Training Objective
Avi Schwarzschild, Max Cembalest, Karthik Rao, Keegan Hines, and John Dickerson

When Fair Classification Meets Noisy Protected Attributes
Avijit Ghosh, Pablo Kvitca, and Christo Wilson

3:00–4:10 **Keynote 2 // Room 510AC**

Changing distributions and preferences in learning systems
Jamie Morgenstern

4:10–4:30 **Short Break**

4:30–6:00 **Proceedings Session 6 // Room 510AC**
Session Chair: Judith Simon

Talks

Disambiguating Algorithmic Bias: From Neutrality to Justice
Elizabeth Edenberg and Alexandra Wood

A sector-based approach to AI ethics: Understanding ethical issues of AI-related incidents within their sectoral context
Dafna Burema, Nicole Debowski-Weimann, Alexander von Janowski, Jil Grabowski, Mihai Maftei, Mattis Jacobs, Patrick van der Smagt, and Djalel Benbouzid

Democratising AI: Multiple Meanings, Goals, and Methods
Elizabeth Seger, Aviv Ovadya, Divya Siddarth, Ben Garfinkel, and Allan Dafoe

Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction
Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk

Student Program Lightning Talks

Exploring the Moral Value of Explainable Artificial Intelligence Through Public Service Postal Banks
Joshua Brand

It Takes A Village To Raise An AI-based System - The Interdisciplinary Work Of Realising Clinical AI in Denmark and Kenya
Hubert Zajac

AI-driven Automation as a Pre-condition for Eudaimonia
Anastasia Siapka

Multi Value Alignment: four steps for aligning ML/AI development choices with multiple values
Hetvi Jethwanii

"Way too good and way beyond comfort": The trade-off between user perception of benefits and comfort in media personalisation
Anna Marie Rezk

Towards formalizing and assessing AI fairness
Anna Schmitz

Is Sortition Both Representative and Fair?
Soroush Ebadian

How and to which extent will the provisions of the Digital Services Act of the European Union impact on the relationship between users and platforms as information providers?
Matteo Fabbri

Lightning Talks

Towards User Guided Actionable Recourse
Jayanth Yetukuri, Ian Hardy, and Yang Liu

Learning from Discriminatory Training Data
Przemyslaw Grabowicz, Nicholas Perello, and Kenta Takatsu

Stress-testing Bias Mitigation Algorithms to Understand Fairness Vulnerabilities
Karan Bhanot, Ioana Baldini, Dennis Wei, Jiaming Zeng, and Kristin Bennett

Perceived Algorithmic Fairness using Organizational Justice Theory: An Empirical Case Study on Algorithmic Hiring
Guusje Juijn, Niya Stoimenova, Joao Reis, and Dong Nguyen

Social Biases through the Text-to-Image Generation Lens
Ranjita Naik and Besmira Nushi

How do you feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection
Philippe Lammerts, Philip Lippmann, Yen-Chia Hsu, Fabio Casati, and Jie Yang

WEDNESDAY, AUGUST 9

6:00–8:00 **Poster Session & Reception // Room 510BD**

Posters

Action Guidance and AI Alignment
Pamela Robinson

Ethical and Social Risks of Generative Text-to-Image Models
Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh

On the Connection between Game-Theoretic Feature Attributions and Counterfactual Explanations
Emanuele Albini, Shubham Sharma, Saumitra Mishra, Danial Dervovic, and Daniele Magazzeni

Adaptive Adversarial Training Does Not Increase Recourse Costs
Ian Hardy, Jayanth Yetukuri, and Yang Liu

REFRESH: Responsible and Efficient Feature Reselection guided by SHAP values
Shubham Sharma, Sanghamitra Dutta, Emanuele Albini, Freddy Lecue, Daniele Magazzeni, and Manuela Veloso

Fairness Implications of Encoding Protected Categorical Attributes
Carlos Mougan, Jose M. Alvarez, Salvatore Ruggieri, and Steffen Staab

Flickr Africa: Examining Geo-Diversity in Large-Scale, Human-Centric Visual Data
Keziah Naggita, Julienne LaChance, and Alice Xiang

Evaluation of targeted dataset collection on racial equity in face recognition
Rachel Hong, Tadayoshi Kohno, and Jamie Morgenstern

Disambiguating Algorithmic Bias: From Neutrality to Justice
Elizabeth Edenberg and Alexandra Wood

Evaluating Biased Attitude Associations of Language Models in an Intersectional Context
Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan

Unmasking Nationality Bias: A Study of Human Perception of Nationalities in AI-Generated Articles
Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson

No Justice, No Robots: From the Dispositions of Policing to an Abolitionist Robotics
Tom Williams and Kerstin Sophie Haring

Why We Need to Know More: Exploring the State of AI Incident Documentation Practices
Violet Turri and Rachel Dzombak

What does it mean to be a responsible AI practitioner: An ontology of roles and skills
Shalaleh Rismani and Ajung Moon

Student Posters

Navigating the Limits of AI Explainability: Designing for Novice Technology Users in Low-Resource Settings
Chinasa Okolo

WEDNESDAY, AUGUST 9

True and Fair: Robust and Unbiased Fake News Detection via Interpretable Machine Learning
Chahat Raj

Benchmarked Ethics: A Roadmap to AI Alignment, Moral Knowledge, and Control
Aidan Kierans

Algorithmic Bias: When stigmatization becomes a perception
Olalekan Akintande

Ethical Principles for Reasoning about Value Preferences
Jessica Woodgate

Explainability in Process Mining: A Framework for Improved Decision-Making
Luca Nannini

Can AlphaGo be apt subjects for Praise/Blame for "Move 37"?
Mubarak Hussain

Anticipatory regulatory instruments for AI systems —
A comparative study of regulatory sandbox schemes
Deborah Morgan

Examining the Ethics of Brain-Computer Interfaces: Ensuring Safety, the Rights and Dignity of Personhood
Terkura Thomas Mchia

Investigating the Relative Strengths of Humans and Machine Learning in Decision-Making
Charvi Rastogi

THURSDAY, AUGUST 10

9:00–9:40 **Proceedings Session 7 // Room 510AC**

Session Chair: Luyao Zhang

Talks

GATE: A Challenge Set for Gender-Ambiguous Translation Examples
Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary

Reclaiming the Digital Commons: A Public Data Trust for Training Data
Alan Chan, Herbie Bradley, and Nitarshan Rajkumar

Lightning Talks

Human Uncertainty in Concept-Based AI Systems
Katherine Collins, Matthew Barker, Mateo Espinosa Zarlenga, Naveen Raman, Umang Bhatt, Mateja Jamnik, Ilia Sucholutsky, Adrian Weller, and Krishnamurthy Dvijotham

ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages
Sourojit Ghosh and Aylin Caliskan

Designing for Human-AI Collaboration in Auditing LLMs with LLMs
Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi

Measures of Disparity and their Efficient Estimation
Harvineet Singh and Rumi Chunara

9:40–11:40 **Poster Session // Room 510BD**

Posters

Reckoning with the Disagreement Problem: Explanation Consensus as a Training Objective
Avi Schwarzschild, Max Cembalest, Karthik Rao, Keegan Hines, and John Dickerson

When Fair Classification Meets Noisy Protected Attributes
Avijit Ghosh, Pablo Kvitca, and Christo Wilson

Towards User Guided Actionable Recourse
Jayanth Yetukuri, Ian Hardy, and Yang Liu

Learning from Discriminatory Training Data
Przemyslaw Grabowicz, Nicholas Perello, and Kenta Takatsu

Stress-testing Bias Mitigation Algorithms to Understand Fairness Vulnerabilities
Karan Bhanot, Ioana Baldini, Dennis Wei, Jiaming Zeng, and Kristin Bennett

Perceived Algorithmic Fairness using Organizational Justice Theory: An Empirical Case Study on Algorithmic Hiring
Guusje Juijn, Niya Stoimenova, Joao Reis, and Dong Nguyen

Social Biases through the Text-to-Image Generation Lens
Ranjita Naik and Besmira Nushi

Evaluating the Fairness of Discriminative Foundation Models in Computer Vision
Junaid Ali, Matthäus Kleindessner, Florian Wenzel, Kailash Budhathoki, Volkan Cevher, and Chris Russell

How do you feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection
Philippe Lammerts, Philip Lippmann, Yen-Chia Hsu, Fabio Casati, and Jie Yang

Human Uncertainty in Concept-Based AI Systems
Katherine Collins, Matthew Barker, Mateo Espinosa Zarlenga, Naveen Raman, Umang Bhatt, Mateja Jamnik, Ilia Sucholutsky, Adrian Weller, and Krishnamurthy Dvijotham

ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages
Sourojit Ghosh and Aylin Caliskan

Designing for Human-AI Collaboration in Auditing LLMs with LLMs
Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi

Measures of Disparity and their Efficient Estimation
Harvineet Singh and Rumi Chunara

Student Posters

Public attitudes toward ethical AI in courts: A Vignette Survey and Deliberation Experiment
Arna Woemmel

How to promote equitable sleep care among people experiencing homelessness: An AI-enabled person-centered computer vision-based solution
Behrad Taghibeyglou

Sealed Knowledges: A Critical Approach to the Usage of LLMs as Search Engines
Nora Freya Lindemann

A Responsible Artificial Intelligence Approach to Epistemic Credibility in Artificial Intelligence Systems
Abiola Azeez

Exploring the Moral Value of Explainable Artificial Intelligence Through Public Service Postal Banks
Joshua Brand

It Takes A Village To Raise An AI-based System - The Interdisciplinary Work Of Realising Clinical AI in Denmark and Kenya
Hubert Zajac

AI-driven Automation as a Pre-condition for Eudaimonia
Anastasia Siapka

Multi Value Alignment: four steps for aligning ML/AI development choices with multiple values
Hetvi Jethwani

"Way too good and way beyond comfort": The trade-off between user perception of benefits and comfort in media personalisation
Anna Marie Rezk

Towards formalizing and assessing AI fairness
Anna Schmitz

Is Sortition Both Representative and Fair?
Soroush Ebadian

How and to which extent will the provisions of the Digital Services Act of the European Union impact on the relationship between users and platforms as information providers?
Matteo Fabbri

Towards a Holistic Approach: Understanding Sociodemographic Biases in NLP Models using an Interdisciplinary Lens
Pranav Narayanan Venkit

Safety in Conversational Systems
Jinhwa Kim

11:40–12:50 **Closing Keynote // Room 510AC**

AI for/by the majority world: From technologies of dispossession to technologies of radical care
Paola Ricaurte Quijano

12:50–1:00 **Conference Closing // Room 510AC**

AIES-23 SPONSORS

AIES 2023 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

Google

Google DeepMind

pwc

SCHMIDT FUTURES

GOLD

future of life INSTITUTE

Sony AI

SILVER

IBM

JPMorgan Chase & Co.

GENEROUS FINANCIAL SUPPORT FOR THE STUDENT PROGRAM PROVIDED BY

NSF

acm SIGAI

ORGANIZING ASSOCIATIONS

Association for the Advancement of Artificial Intelligence

acm

acm SIGAI