The Second AAAI / ACM Annual Conference on AI, Ethics, and Society

January 26 - January 28

Hilton Hawaiian Village
Honolulu, Hawaii, USA

Program Guide

AIES-19 Conference Overview

On Saturday Jan. 26 at 7:30pm, we will have a **panel on "Perspectives on Responsible AI"** with Peter Hershock (moderator), Yolanda Gil, Huw Price, Francesca Rossi, and Dekai Wu. Location: *Spalding Building, Room 155, University of Hawaii - Manoa.* **The main program** (Jan. 27-28) takes place at the *Mid Pacific Conference Center at the Hilton Hawaiian Village, Coral 4*.

	Sunday, January 27	**Monday, January 28**
8:40 - 8:50	Opening remarks	
8:50 - 9:40	Invited talk **Ryan Calo (**Univ. of Washington)	Invited talk **Anca Dragan** (UC Berkeley)
9:40 - 10:00	Spotlight 1: *Normative Perspectives*	Spotlight 3: *Empirical Perspectives*
10:00 - 10:30	Coffee + Posters	Coffee + Posters
10:30 - 11:30	Session 1: *Algorithmic Fairness*	Session 6: *Social Science Models*
11:30 - 12:30	Session 2: *Norms and Explanations*	Session 7: *Measurement & Justice*
12:30 - 2:00	Lunch	Lunch
2:00 - 3:00	Session 3: *Artificial Agency*	Session 8: *AI for Social Good*
3:00 - 4:00	Session 4: *Autonomy & Lethality*	Session 9: *Human-Machine Interaction*
4:00 - 4:30	Coffee + Posters	Coffee + Posters
4:30 - 5:30	Session 5: *Rights & Principles* Spotlight 2: *Fairness+Explanations*	Invited talk **David Danks (**Carnegie Mellon) Closing remarks
5:40 - 6:30	Invited talk **Susan Athey** (Stanford University)	
7:00 - 9:00	Reception + Posters	

Acknowledgments

Steve Mckinlay (Wellington Institute of Technology)

Kuldeep S. Meel (National Univ. of Singapore)

Jason Millar (Univ. of Ottawa)

Ted Parson (UCLA)

Dominik Peters (Univ. of Oxford)

Huw Price (Univ. of Cambridge)

Andrea Renda (CEPS)

Juan Antonio Rodriguez Aguilar (IIIA-CSIC)

Heather Roff (Johns Hopkins Univ.)

Francesca Rossi (IBM)

Michael Rovatsos (The Univ. of Edinburgh)

Stuart Russell (UC Berkeley)

Jana Schaich Borg (Duke Univ.)

Matthias Scheutz (Tufts Univ.)

Susan Schneider (Univ. of Connecticut)

Munindar Singh (North Carolina State Univ.)

Walter Sinnott-Armstrong (Duke Univ.)

Joshua Skorburg (Duke Univ.)

Marija Slavkovik (Univ. of Bergen)

Eric Sodomka (Facebook)

John Sullins (Sonoma State Univ.)

Kristen Brent Venable (Tulane Univ. and IHMC)

Toby Walsh (The Univ. of New South Wales)

Robin Zebrowski (Beloit College)

Ben Zevenbergen (Princeton Univ.)

Student Program Chairs:

Kristen Brent, Venable (Tulane Univ. and IHMC)

Emanuelle Burton (Univ. of Illinois at Chicago)

Student Program Committee:

Lindsey Barrett (Georgetown Univ.)

Andrea Loreggia (Univ. of Padova)

Nicholas Mattei (IBM)

John P. Dickerson (Univ. of Maryland)

Dan Estrada (Rutgers)

Judy Goldsmith (Univ. of Kentucky)

Michael Rovatsos (The Univ. of Edinburgh)

Auxiliary Reviewers

Benjamin Abramowitz (Rensselaer Polytechnic Institute)

Adel Abusittac (Polytechnique Montréal)

Nirav Ajmeri (North Carolina State University)

Joe Austerweil (Univ. of Wisconsin)

William Bailey (Univ. of Kentucky)

Avinash Balakrishnan (IBM Research)

Yahav Bechavod (Hebrew Univ.)

Debarun Bhattacharjya (IBM)

Malik Boykin (Brown Univ.)

Gordon Briggs (Naval Research Laboratory)

Mithun Chakraborty (National Univ. of Singapore)

Meia Chita-Tegmark (Tufts Univ.)

Liron Cohen (Univ. of Southern California)

Cristina Cornelio (IBM Research)

Alan Davoust (Univ. of Edinburgh)

Upol Ehsan (Cornell Univ.)

Olivia Johanna Erdelyi (Univ. of Canterbury)

Bohui Fang (Jiao Tong Univ.)

Rupert Freeman (Microsoft)

Anne-Marie George (Technische Universität Berlin)

Paul Gölz (Carnegie Mellon Univ.)

Sriram Gopalakrishnan (Arizona State Univ.)

Sachin Grover (Arizona State Univ.)

Peter Haas (Brown Univ.)

Chung-Wei Hang (IBM)

Kerstin Sophie Haring (United States Air Force Academy)

Brent Harrison (Univ. of Kentucky)

Jobst Heitzig (Potsdam Institute for Climate Impact Research)

Matthew Joseph (Univ. of Pennsylvania)

Dan Kasenberg (Tufts Univ.)

Anagha Kulkarni (Arizona State Univ.)

Barton Lee (Univ. of New South Wales Sydney)

Jiaoyang Li (Univ. of Southern California)

Maite Lopez-Sanchez (Univ. of Barcelona)

Hang Ma (Univ. of Southern California)

Pradeep Kumar Murukannaiah (Rochester Institute of Technology)

Kamran Najeebullah (CSIRO)

Seth Neel (Univ. of Pennsylvania)

Hoon Oh (Carnegie Mellon Univ.)

Manish Raghavan (Cornell Univ.)

David Robinson (Cornell Univ.)

Abdallah Saffidine (Australian National Univ.)

Candice Schumann (Univ. of Maryland)

Andrew Selbst (Yale Information Society Project)

Sailik Sengupta (Arizona State Univ.)

Nolan Shaw (Univ. of Waterloo)

Cory Siler (Univ. of Kentucky)

Jakub Sliwinski (ETH Zürich)

Sarath Sreedharan (Arizona State Univ.)

Ana-Andreea Stoica (Columbia Univ.)

Zhaohong Sun (Univ. of New South Wales Sydney)

Tomas Trescak (Western Sydney Univ.)

Tansel Uras (Univ. of Southern California)

Kush Varshne (IBM Research)

Ellen Vitercik (Carnegie Mellon Univ.)

John Voiklis (New Knowledge Organization, LTD)

Patrick Watson (IBM Research)

Hong Xu (NHLBI)

Guangchao Yuan (Microsoft)

Yantian Zha (Arizona State Univ.)

Zhe Zhang (IBM)

Yair Zick (National Univ. of Singapore)

Website: Francisco Cruz (IIIA)

Conference Assistant: Kenzie Doyle (Duke University)

Registration

Onsite registration and badge pick-up will be located in the foyer of the Coral Ballroom (outside Coral 2) on the sixth floor of the Mid Pacific Conference Center at the Hilton Hawaiian Village, 2005 Kalia Road, Honolulu, HI 96815, USA. Registration hours will be Sunday and Monday, 7:30 AM – 5:00 PM. All attendees must pick up their registration packets/badges for admittance to programs.

Wifi   Select Hilton Meetings Password: AAAI19

Social Events

Student lunch, Jan 28 during the lunch break 12:30PM-2PM (only for those in the student program)

Where: MW Restaurant, 1538 Kapiolani Blvd #107, Honolulu, HI 96814

20 minute walk or 8 minute ride from the conference venue.

AIES 2019 Program

January 26th

Responsible Artificial Intelligence: Aligning Technology, Engineering and Ethics
7:30-9:00pm Saturday, January 26, 2019
Spalding Building, Room 155, University of Hawaii - Manoa

This moderated panel discussion will focus on the societal implications of progress in artificial intelligence and how perspectives on this vary around the world. The panel is open to all and no formal background in AI is required. The panelists include:

· **Dr. Yolanda Gil,** Professor of Computer Sciences and Spatial Sciences at the University of Southern California, Director of the USC Center for Knowledge-Powered Interdisciplinary Data Science, and the 24th President of the Association for the Advancement of Artificial Intelligence
· **Dr. Huw Price**, Professor of Philosophy at University of Cambridge and Academic Director of its Centre for the Study of Existential Risk and the Leverhulme Centre for the Future of Intelligence
· **Dr. Francesca Rossi**, the IBM AI Ethics Global Leader, distinguished research scientist at the IBM T.J. Watson Research Center, and Professor of Computer Science at the University of Padova
· **Dr. Wu Dekai,** Professor of Computer Science and Engineering at Hong Kong University of Science and Technology and a founding fellow of the Association for Computational Linguistics

The panel will be moderated by **Dr. Peter Hershock**, Director of the Asian Studies Development Program at the East-West Center.

The main program (Jan. 27-28, everything below) is at the *Mid Pacific Conference Center at the Hilton Hawaiian Village,* in **Coral 4** (sessions, some posters on the 28th), **Coral 3** (reception, posters on the 27th) and foyer (some posters on the 28th).

January 27th

8:40 AM - 8:50 AM   **Opening Remarks**

8:50 AM - 9:40 AM   **Invited Talk (**Chair: Vincent Conitzer)

Ryan Calo (Univ. of Washington School of Law):   *How We Talk About AI (And Why It Matters)*

Abstract: How we talk about artificial intelligence matters. Not only do our rhetorical choices influence public expectations of AI, they implicitly make the case for or against specific government interventions. Conceiving of AI as a global project to which each nation can contribute, for instance, suggests a different course of action than understanding AI as a "race" America cannot afford to lose. And just as inflammatory terms such as "killer robot" aim to catalyze limitations of autonomous weapons, so do the popular terms "ethics" and "governance" subtly argue for a lesser role for government in setting AI policy. How should we talk about AI? And what's at stake with our rhetorical choices? This presentation explores the interplay between claims about AI and law's capacity to channel AI in the public interest.

9:40 AM - 10:00 AM *Spotlight 1: Normative Perspectives* (Chair: Vincent Conitzer)

Rightful Machines and Dilemmas
Ava Thomas Wright

Modelling and Influencing the AI Bidding War: A Research Agenda
The Anh Han, Luis Moniz Pereira and Tom Lenaerts

The Heart of the Matter: Patient Autonomy as a Model for the Wellbeing of Technology Users
Emanuelle Burton, Kristel Clayville, Judy Goldsmith and Nicholas Mattei

Requirements for an Artificial Agent with Norm Competence
Bertram Malle, Paul Bello and Matthias Scheutz

Toward the Engineering of Virtuous Machines
Naveen Sundar Govindarajulu, Selmer Bringsjord and Rikhiya Ghosh

Semantics Derived Automatically from Language Corpora Contain Human-like Moral Choices
Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf and Kristian Kersting

Ethically Aligned Opportunistic Scheduling for Productive Laziness
Han Yu, Chunyan Miao, Yongqing Zheng, Lizhen Cui, Simon Fauvel and Cyril Leung

(When) Can AI Bots Lie?
Tathagata Chakraborti and Subbarao Kambhampati

Epistemic Therapy for Bias in Automated Decision-Making
Thomas Gilbert and Yonatan Mintz

Algorithmic greenlining: An approach to increase diversity
Christian Borgs, Jennifer Chayes, Nika Haghtalab, Adam Kalai and Ellen Vitercik

10:00 AM - 10:30 AM **Coffee Break + Poster Session 1 (Posters from *Spotlight 1*)**

10:30 AM - 11:30 AM *Session 1: Algorithmic Fairness* (Chair: Gillian Hadfield)

Active Fairness in Algorithmic Decision Making
Michiel Bakker, Alejandro Noriega-Campero, Bernardo Garcia-Bulle and Alex Pentland

Paradoxes in Fair Computer-Aided Decision Making
Andrew Morgan and Rafael Pass

Fair Transfer Learning with Missing Protected Attributes
Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush Varshney, Skyler Speakman, Zairah Mustahsan and Supriyo Chakraborty

How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness
Nripsuta Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David Parkes and Yang Liu

11:30 AM - 12:30 PM *Session 2: Norms and Explanations* (Chair: Vincent Conitzer)

Learning Existing Social Conventions via Observationally Augmented Self-Play *(Best Paper Award)*
Alexander Peysakhovich and Adam Lerer

Legible Normativity for AI Alignment: The Value of Silly Rules
Dylan Hadfield-Menell, Mckane Andrus and Gillian Hadfield

TED: Teaching AI to Explain its Decisions
Noel Codella, Michael Hind, Karthikeyan Natesan Ramamurthy, Murray Campbell, Amit Dhurandhar, Kush Varshney, Dennis Wei and Aleksandra Mojsilovic

Understanding Black Box Model Behavior through Subspace Explanations
Himabindu Lakkaraju, Ece Kamar, Rich Caruana and Jure Leskovec

12:30 PM - 2:00 PM   **Lunch Break**

2:00 PM - 3:00 PM   ***Session 3: Artificial Agency*** (Chair: Shannon Vallor)

Shared Moral Foundations of Embodied Artificial Intelligence*
    Joe Cruz

Building Jiminy Cricket: An Architecture for Moral Agreements Among Stakeholders
    Beishui Liao, Marija Slavkovik and Leendert van der Torre

AI + Art = Human*
    Antonio Daniele and Yi-Zhe Song

Speaking on Behalf: Representation, Delegation, and Authority in Computational Text Analysis
    Eric Baumer and Micki McGee

3:00 PM - 4:00 PM   ***Session 4: Autonomy and Lethality*** (Chair: Ryan Calo)

Killer Robots and Human Dignity*
    Daniel Lim

Regulating Lethal and Harmful Autonomy: Drafting a Protocol VI of the Convention on Certain Conventional Weapons*
    Sean Welsh

Balancing the Benefits of Autonomous Vehicles
    Timothy Geary and David Danks

Compensation at the Crossroads: Autonomous Vehicles and Alternative Victim Compensation Schemes
    Tracy Pearl

4:00 PM - 4:30 PM   **Coffee Break + Student Poster Session**

4:30 PM - 5:00 PM   ***Session 5: Rights and Principles*** (Chair: Shannon Vallor)

The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions
    Jess Whittlestone, Rune Nyrup, Anna Alexandrova and Stephen Cave

How Technological Advances Can Reveal Rights
    Jack Parker and David Danks

5:00 PM - 5:30 PM   ***Spotlight 2: Fairness and Explanations*** (Chair: Shannon Vallor)

IMLI: An Incremental Framework for MaxSAT-Based Learning of Interpretable Classification Rules
    Bishwamittra Ghosh and Kuldeep S. Meel

Loss-Aversively Fair Classification
    Junaid Ali, Muhammad Bilal Zafar, Adish Singla and Krishna P. Gummadi

Counterfactual Fairness in Text Classification through Robustness
    Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed Chi and Alex Beutel

Taking Advantage of Multitask Learning for Fair Classification
    Luca Oneto, Michele Doninini, Amon Elders and Massimiliano Pontil

Explanatory Interactive Machine Learning
    Stefano Teso and Kristian Kersting

Multiaccuracy: Black-Box Post-Processing for Fairness in Classification
    Michael P. Kim, Amirata Ghorbani and James Zou

A Formal Approach to Explainability
    Lior Wolf, Tomer Galanti and Tamir Hazan

Costs and Benefits of Fair Representation Learning
> Daniel McNamara, Cheng Soon Ong and Robert Williamson

Creating Fair Models of Atherosclerotic Cardiovascular Disease Risk
> Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan and Nigam Shah

Global Explanations of Neural Networks: Mapping the Landscape of Predictions
> Mark Ibrahim, Melissa Louie, Ceena Modarres and John Paisley

Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure
> Alexander Amini, Ava Soleimany, Wilko Schwarting, Sangeeta Bhatia and Daniela Rus

Crowdsourcing with Fairness, Diversity and Budget Constraints
> Naman Goel and Boi Faltings

What are the biases in my word embedding?
> Nathaniel Swinger, Maria De-Arteaga, Neil Heffernan Iv, Mark Dm Leiserson and Adam Kalai

Equalized Odds Implies Partially Equalized Outcomes Under Realistic Assumptions
> Daniel McNamara

The Right To Confront Your Accuser: Opening the Black Box of Forensic DNA Software
> Jeanna Matthews, Marzieh Babaeianjelodar, Stephen Lorenz, Abigail Matthews, Mariama Njie, Nathan Adams, Dan Krane, Jessica Goldthwaite and Clinton Hughes

5:30 PM - 5:40 PM   **Bio Break / Slack**

5:40 PM - 6:30 PM   **Invited Talk** (Chair: Gillian Hadfield)

Susan Athey (Stanford University, Graduate School of Business and Economics):

Guiding and Implementing AI

Abstract: This talk will provide theoretical perspectives on how organizations should guide and implement AI in a way that is fair and that achieves the organization's objectives. We first consider the role of insights from statistics and causal inference for this problem. We then extend the framework to incorporate considerations about designing the rewards for AI and human decision-makers, and designing tasks and authority to optimally combine AI and humans to achieve the most effective incentives.

7:00 PM -- 9:00 PM   **Reception + Poster Session 2 (Posters from *Spotlight 2, Session 1, and Session 2*)**
(with exception of papers marked *)

January 28th

8:50 AM - 9:40 AM   **Invited Talk** (Chair: Vincent Conitzer)

Anca Dragan (UC Berkeley, EECS): *Specifying AI Objectives as a Human-AI Collaboration Problem*

Abstract: Estimation, planning, control, and learning are giving us robots that can generate good behavior given a specified objective and set of constraints. What I care about is how humans enter this behavior generation picture, and study two complementary challenges: 1) *how* to optimize behavior when the robot is not acting in isolation, but needs to coordinate or collaborate with people; and 2) *what* to optimize in order to get the behavior we want. My work has traditionally focused on the former, but more recently I have been casting the latter as a human-robot collaboration problem as well (where the human is the end-user, or even the robotics engineer building the system). Treating it as such has enabled us to use robot actions to gain information; to account for human pedagogic behavior; and to exchange information between the human and the robot via a plethora of communication channels, from external forces that the person physically applies to the robot, to comparison queries, to defining a proxy objective function.

9:40 AM - 10:00 AM *Spotlight 3: Empirical Perspectives* (Chair: Gillian Hadfield)

"Scary Robots": Examining Public Responses to AI
> Stephen Cave, Kate Coughlan and Kanta Dihal

Framing Artificial Intelligence in American Newspaper
> Ching-Hua Chuan, Wan-Hsiu Tsai and Su Yeon Cho

Perceptions of Domestic Robots' Normative Behavior Across Cultures
> Huao Li, Stephanie Milani, Vigneshram Krishnamoorthy, Michael Lewis and Katia Sycara

Mapping Missing Population in Rural India: A Deep Learning Approach with Satellite Imagery
> Wenjie Hu, Jay Harshadbhai Patel, Zoe-Alanah Robert, Paul Novosad, Samuel Asher, Zhongyi Tang, Marshall Burke, David Lobell and Stefano Ermon

Mapping Informal Settlements in Developing Countries using Machine Learning and Low Resolution Multi-spectral Data
> Bradley Gram-Hansen, Patrick Helber, Indhu Varatharajan, Faiza Azam, Alejandro Coca-Castro, Veronika Kopackova and Piotr Bilinski

Human-AI Learning Performance in Multi-Armed Bandits
> Ravi Pandya, Sandy Huang, Dylan Hadfield-Menell and Anca Dragan

A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity
> De'Aira Bryant and Ayanna Howard

Degenerate Feedback Loops in Recommender Systems
> Ray Jiang, Silvia Chiappa, Tor Lattimore, Andras Gyorgy and Pushmeet Kohli

TrolleyMod v1.0: An Open-Source Simulation and Data-Collection Platform for Ethical Decision Making in Autonomous Vehicles
> Vahid Behzadan, James Minton and Arslan Munir

The Seductive Allure of Artificial Intelligence-Powered Neurotechnology
> Charles Giattino, Lydia Kwong, Chad Rafetto and Nita Farahany

10:00 AM - 10:30 AM **Coffee Break + Poster Session 3 (Posters from *Spotlight 3, Sessions 3-5*)**
> (with exception of papers marked *)

10:30 AM - 11:30 PM *Session 6: Social Science Models for AI* (Chair: Susan Athey)

Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures*
> Daniel Susser

Reinforcement learning and inverse reinforcement learning with system 1 and system 2
> Alexander Peysakhovich

Incomplete Contracting and AI Alignment
> Dylan Hadfield-Menell and Gillian Hadfield

Theories of parenting and their application to artificial intelligence*
> Sky Croeser and Peter Eckersley

11:30 AM - 12:30 PM *Session 7: Measurement and Justice* (Chair: David Danks)

Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products (Best Student Paper Award)
> Inioluwa Deborah Raji and Joy Buolamwini

A framework for benchmarking discrimination-aware models in machine learning
> Rodrigo L. Cardoso, Wagner Meira Jr., Virgilio Almeida and Mohammed J. Zaki

Towards a Just Theory of Measurement: A Principled Social Measurement Assurance Program*
> Thomas Gilbert and McKane Andrus

Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements
    Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof and Ed H. Chi

12:30 PM - 2:00 PM   **Lunch Break**

2:00 PM - 3:00 PM   **Session 8: AI for Social Good** (Chair: Vincent Conitzer)

On Influencing Individual Behavior for Reducing Transportation Energy Expenditure in a Large Population
    Shiwali Mohan, Frances Yan, Victoria Bellotti, Ahmed Elbery, Hesham Rakha and Matthew Klenk

Guiding Prosecutorial Decisions with an Interpretable Statistical Model
    Zhiyuan Lin, Alex Chohlas-Wood and Sharad Goel

Using deceased-donor kidneys to initiate chains of living donor kidney paired donations: algorithm and experimentation
    Cristina Cornelio, Lucrezia Furian, Antonio Nicolò and Francesca Rossi

Inferring Work Task Automatability from AI Expert Evidence
    Paul Duckworth, Logan Graham and Michael Osborne

3:00 PM - 4:00 PM   *Session 9: Human and Machine Interaction* (Chair: Esma Aimeur)

Robots Can Be More Than Black And White: Examining Racial Bias Towards Robots
    Arifah Addison, Kumar Yogeeswaran and Christoph Bartneck

Tact in Noncompliance: The Need for Pragmatically Apt Responses to Unethical Commands
    Ryan Blake Jackson, Ruchen Wen and Tom Williams

AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI*
    Karina Vold and Jose Hernandez-Orallo

Human Trust Measurement Using an Immersive Virtual Reality Autonomous Vehicle Simulator
    Shervin Shahrdar, Corey Park and Mehrdad Nojoumian

4:00 PM - 4:30 PM   *Coffee Break + Poster Session 4* (Posters from *Sessions 6-9*)
    (with exception of papers marked *)

4:30 PM - 5:20 PM   **Invited Talk** (Chair: Shannon Vallor)

David Danks (Carnegie-Mellon University, Dept. of Philosophy)   *The Value of Trustworthy AI*

Abstract: There are an increasing number of calls for "AI that we can trust," but rarely with any clarity about what 'trustworthy' means or what kind of value it provides. At the same time, trust has become an increasingly important and visible topic of research in AI, HCI, and HRI communities. In this talk, I will first unpack the notion of 'trustworthy', from both philosophical and psychological perspectives, as it might apply to an AI system. In particular, I will argue that there are different kinds of (relevant, appropriate) trustworthiness, depending on one's goals and modes of interaction with the AI. There is not just one kind of trustworthy AI, even though trustworthiness (of the appropriate type) is arguably the primary feature that we should want in an AI system. Trustworthiness is both more complex, and also more important, than standardly recognized in the public calls-to-action (and this analysis connects and contrasts in interesting ways with others).

5:20 - 5:30   **Closing Remarks**

Student Program

This year 23 exceptional students from different disciplines including computer science, political science, and law were selected to participate in the conference. These students received a free registration and a travel grant. Moreover, the accepted students will:

·   present a poster (during the student poster session, in the afternoon of Jan 27th),

·   meet senior members of the community,

·   be invited to the student lunch on Jan 28th,

·   have the opportunity to publish a 2-page extended abstract of their work in the proceedings.

Awards (sponsored by the Partnership on AI)

Best paper award: Alexander Peysakhovich and Adam Lerer: **Learning Existing Social Conventions via Observationally Augmented Self-Play**

Best student paper award: Inioluwa Deborah Raji and Joy Buolamwini: **Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products**

Invited Speakers

Susan Athey is The Economics of Technology Professor at Stanford Graduate School of Business. She received her bachelor's degree from Duke University and her Ph.D. from Stanford, and she holds an honorary doctorate from Duke University. She previously taught at the economics departments at MIT, Stanford and Harvard. In 2007, Professor Athey received the John Bates Clark Medal, awarded by the American Economic Association to "that American economist under the age of forty who is adjudged to have made the most significant contribution to economic thought and knowledge." She was elected to the National Academy of Science in 2012 and to the American Academy of Arts and Sciences in 2008. Professor Athey's research focuses on the intersection of machine learning and econometrics, marketplace design, and the economics of digitization. She advises governments and businesses on marketplace design and platform economics, serving as consulting chief economist to Microsoft for a number of years, and serving on the boards of Expedia, Lending Club, Rover, Ripple, and Turo.

Ryan Calo is the Lane Powell and D. Wayne Gittinger Associate Professor at the University of Washington School of Law. He is a faculty co-director (with Batya Friedman and Tadayoshi Kohno) of the University of Washington Tech Policy Lab, a unique, interdisciplinary research unit that spans the School of Law, Information School, and Paul G. Allen School of Computer Science and Engineering. Professor Calo holds courtesy appointments at the University of Washington Information School and the Oregon State University School of Mechanical, Industrial, and Manufacturing Engineering.

David Danks is the L.L. Thurstone Professor of Philosophy & Psychology, and Head of the Department of Philosophy, at Carnegie Mellon University. He is also an adjunct member of the Heinz College of Information Systems and Public Policy, and the Center for the Neural Basis of Cognition. His research interests are at the intersection of philosophy, cognitive science, and machine learning, using ideas, methods, and frameworks from each to advance our understanding of complex, interdisciplinary problems. In particular, Danks has examined the ethical, psychological, and policy issues around AI and robotics in transportation, healthcare, privacy, and security. He has received a McDonnell Foundation Scholar Award, an Andrew Carnegie Fellowship, and funding from multiple agencies.

Anca Dragan is an Assistant Professor in EECS at UC Berkeley, where she runs the InterACT lab. Her goal is to enable robots to work with, around, and in support of people. Anca did her PhD in the Robotics Institute at Carnegie Mellon University on legible motion planning. At Berkeley, she helped found the Berkeley AI Research Lab, is a co-PI for the Center for Human-Compatible AI, and has been honored by the Sloan fellowship, the NSF CAREER award, the Okawa award, MIT's TR35, and an IJCAI Early Career Spotlight.

Organizing Associations

AIES 2019 would like to thank the organizing associations that shared the need, the opportunity, and the vision to start a new conference series on the topics of AI, ethics, and society. They wholeheartedly supported the program chairs and all others involved in the organization with resources, advice, connections, and organizational support.

AAAI   aaai.org

Founded in 1979, the Association for the Advancement of Artificial Intelligence (AAAI) is a nonprofit scientific society devoted to advancing the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines. AAAI aims to promote research in, and responsible use of, artificial intelligence. AAAI also aims to increase public understanding of artificial intelligence, improve the teaching and training of AI practitioners, and provide guidance for research planners and funders concerning the importance and potential of current AI developments and future directions.

ACM   https://www.acm.org/

ACM, the Association for Computing Machinery, is the world's largest educational and scientific computing society, uniting educators, researchers and professionals to inspire dialogue, share resources and address the field's challenges. ACM strengthens the computing profession's collective voice through strong leadership, promotion of the highest standards, and recognition of technical excellence. ACM supports the professional growth of its members by providing opportunities for life-long learning, career development, and professional networking.

ACM Special Interest Group on Artificial Intelligence   https://sigai.acm.org/

Who we are: academic and industrial researchers, practitioners, software developers, end users, and students.

What we do:

• Promote and support the growth and application of AI principles and techniques throughout computing

• Sponsor or co-sponsor high-quality, AI-related conferences

• Publish the quarterly newsletter AI Matters and its namesake blog

• Organize the Career Network and Conference (SIGAI CNC) for early-stage researchers in AI

• Sponsor recognized AI awards

• Support important journals in the field

• Provide scholarships to student members to attend conferences

• Promote AI education and publications through various forums and the ACM digital library

Sponsors

Berkeley Existential Risk Initiative   http://existence.org

BERI is a 501(c)3 nonprofit whose mission is to improve human civilization's long-term prospects for survival and flourishing. Its main strategy is to identify technologies that may pose significant civilization-scale risks, and to promote and provide support for research and other activities aimed at reducing those risks.

DeepMInd Ethics & Society   https://deepmind.com/applied/deepmind-ethics-society/

We created DeepMind Ethics & Society because we believe AI can be of extraordinary benefit to the world, but only if held to the highest ethical standards. Technology is not value neutral, and technologists must take responsibility for the ethical and social impact of their work. In a field as complex as AI this is easier said than done, which is why we are committed to deep research into ethical and social questions, the inclusion of many voices, and ongoing critical reflection.

Google https://www.google.com/

Google's mission is to organize the world's information and make it universally accessible and useful. AI is helping us do that in exciting new ways, solving problems for our users, our customers, and the world. AI is making it easier for people to do things every day, whether it's searching for photos of loved ones, breaking down language barriers in Google Translate, typing emails on the go, or getting things done with the Google Assistant. AI also provides new ways of looking at existing problems, from rethinking healthcare to advancing scientific discovery. And most importantly of all, we think AI will have the greatest impact when everyone can access it, and when it's built with everyone's benefit in mind. Core to this approach is publishing our research in a wide variety of venues, and open sourcing our tools and systems, like TensorFlow. And while we fundamentally believe in the promise of AI, we also think that it's critical that this technology is used to help people — that it is socially beneficial, fair, accountable, and works for everyone. Learn more about our work to apply AI to the world's most pressing problems, and about the Google AI Impact Challenge, our initiative to support organizations using AI for social good.   Visit https://ai.google/about/

IBM Research AI   https://www.research.ibm.com/artificial-intelligence/

IBM Research is one of the world's largest and most influential corporate research labs, driving research, discovery and innovation from its 12 labs located across six continents. From artificial intelligence breakthroughs in speech and vision technologies, to advancing trust and transparency in algorithms, our IBM Research AI team is pioneering the most promising and disruptive technologies that will transform industries and society.

Amazon   https://www.amazon.com/

Amazon strives to be Earth's most customer-centric company where people can find and discover virtually anything they want to buy online. The world's brightest technology minds come to Amazon to research and develop technology that improves the lives of shoppers, sellers and developers. Building AI that is fair and unbiased is an important pillar of Amazon's thematic framework. As partners in the Partnership for AI, Amazon is continuing to find ways to support fundamental and applied research focused on fair AI to ensure that AI systems demonstrate the fairness, transparency,

explainability, accountability, and preservation of privacy. For more information about Research at Amazon, visit: amazon.jobs/AAAI. Contact Info: AAAI2019@amazon.com

Facebook https://research.fb.com/

At Facebook, ensuring the responsible and thoughtful use of AI is foundational to everything we do—from the data labels we use, to the individual algorithms we build, to the systems they are a part of. To help do this, we're developing new tools like Fairness Flow, which can help generate metrics for evaluating whether there are unintended biases in certain models.

Future of Life Institute https://futureoflife.org/

We are a charity and outreach organization working to ensure that tomorrow's most powerful technologies are beneficial for humanity. With less powerful technologies such as fire, we learned to minimize risks largely by learning from mistakes. With more powerful technologies such as nuclear weapons, synthetic biology and future strong artificial intelligence, planning ahead is a better strategy than learning from mistakes, so we support research and other efforts aimed at avoiding problems in the first place. We are currently focusing on keeping artificial intelligence beneficial and we are also exploring ways of reducing risks from nuclear weapons and biotechnology.

National Science Foundation https://www.nsf.gov/

The National Science Foundation (NSF) is an independent federal agency created by Congress in 1950 "to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense..." NSF is vital because we support basic research and people to create knowledge that transforms the future. This type of support: Is a primary driver of the U.S. economy. Enhances the nation's security. Advances knowledge to sustain global leadership. With an annual budget of $7.5 billion (FY 2017), we are the funding source for approximately 24 percent of all federally supported basic research conducted by America's colleges and universities. In many fields such as mathematics, computer science and the social sciences, NSF is the major source of federal backing.

PricewaterhouseCoopers https://www.pwc.com/

PwC New Services and Emerging Technologies Group works with clients to identify new opportunities and realize value from emerging technologies including Artificial Intelligence and Robotics. We investigate topics across all aspects of the analytics and artificial intelligence pipeline including: collecting, processing, modeling, and productionizing data solutions, with deep specialization in machine learning, deep learning, natural language, simulation and working with data at scale. PwC is also actively working on Responsible AI (building interpretable, ethical, explainable, robust and safe AI) and National AI Strategies.

The Partnership on AI https://www.partnershiponai.org/

The University of Hawaiʻi at Mānoa https://manoa.hawaii.edu/

The University of Hawaiʻi at Mānoa was founded as a small college in 1907. The college became the University of Hawaiʻi in 1920 with the addition of a College of Arts and Sciences.The university continued to grow throughout the 1930s. The Oriental Institute, forerunner of the East-West Center, was founded in 1935, bolstering the university's mounting prominence in Asia-Pacific studies. The university continued to expand throughout the second half of the century and in 1972 was renamed the University of Hawaiʻi at Mānoa to distinguish it from the other campuses in the growing University of Hawaiʻi System. UH Mānoa's School of Law opened in temporary buildings in 1973. The Center for Hawaiian Studies was established in 1977 followed by the School of Architecture in 1980. The School of Ocean and Earth Sciences and Technology was founded eight years later and in 2005 the John A. Burns School of Medicine moved to its present location in Honolulu's Kakaʻako district. Today UH Mānoa is a research university of international standing offering a comprehensive array of undergraduate, graduate, and professional degrees.

Organizing associations:

Sponsors:

Gold level:

Silver level:

Bronze level:

facebook

amazon

pwc

future of life INSTITUTE

Other contributions:

PARTNERSHIP ON AI

UNIVERSITY OF HAWAI'I · MĀLAMALAMA · UA MAU KE EA O KA 'ĀINA I KA PONO · 1907

Alphabetical Speaker/Author Index

Nathan Adams, Forensic Bioinformatic Services
Arifah Addison, University of Canterbury
Anna Alexandrova, Leverhulme Centre for the FOI
Junaid Ali, Max Planck Institute for Software Systems
Virgilio Almeida, Universidade Federal de Minas Gerais
Alexander Amini, MIT
McKane Andrus, UC Berkeley
Samuel Asher, World Bank
Susan Athey, Stanford University
Faiza Azam, University of Bremen
Marzieh Babaeianjelodar, Clarkson University
Michiel Bakker, MIT
Christoph Bartneck, University of Canterbury
Eric P.S. Baumer, Lehigh University
Vahid Behzadan, Kansas State University
Paul Bello, US Naval Research Laboratory

Victoria Bellotti, Palo Alto Research Center
Alex Beutel, Google AI
Sangeeta Bhatia, MIT
Piotr Bilinski, University of Oxford
Jonathan Bischof, Google
Christian Borgs, Microsoft Research
Selmer Bringsjord, Rensselaer Polytechnic Institute
De-Aira Bryant, Georgia Institute of Technology
Joy Buolamwini, MIT
Marshall Burke, Stanford University
Emanuelle Burton, University of Illinois at Chicago
Ryan Calo, University of Washington School of Law
Murray Campbell, IBM
Rodrigo L. Cardoso, Univ. Federal de Minas Gerais
Rich Caruana, Microsoft Research
Stephen Cave, Leverhulme Centre for the FOI
Tathagata Chakraborti, IBM Research AI

Supriyo Chakraborty, IBM
Jennifer Chayes, Microsoft Research
Jilin Chen, Google
Ed H. Chi, Google AI
Silvia Chiappa, DeepMind
Su Yeon Cho, University of Miami
Alex Chohlas-Wood, Stanford University
Ching-Hua Chuan,, University of Miami
Kristel Clayville, Eureka College
Alejandro Coca-Castro, Kings College London
Noel Codella, MIT
Vincent Conitzer, Duke University
Cristina Cornelio, IBM
Amanda Coston, IBM Research/Carnegie Mellon Univ.
Kate Coughlan, BBC
Adrien Coulet, Stanford University
Sky Croeser, Curtin University
Joe Cruz, Williams College
Lizhen Cui, Shandong University
Antonio Daniele, EECS-Queen Mary Univ. of London
David Danks, Carnegie Mellon University
Maria De-Arteaga, Carnegie Mellon University
Evan DeFilippis, Harvard University
Amit Dhurandhar, IBM
Kanta Dihal, Leverhulme Centre for the FOI
Michele Doninini, Amazon
Tulsee Doshi, Google
Anca Dragan, UC Berkeley
Paul Duckworth, University of Oxford
Peter Eckersley, Partnership on AI
Ahmed Elbery, Virginia Tech Transportation Institute
Amon Elders, IIT
Stefano Ermon, Stanford University
Boi Faltings, Artificial Intelligence Lab, EPFL
Nita Farahany, Duke University
Simon Fauvel, Nanyang Technological University
Lucrezia Furian, University of Padova
Tomer Galanti, Tel-Aviv University
Bernardo Garcia-Bulle, MIT
Sahaj Garg, Stanford University
Timothy Geary, CSU Monterey Bay
Amirata Ghorbani, Stanford University
Rikhiya Ghosh, Rensselaer Polytechnic Institute
Bishwamittra Ghosh, National University of Singapore
Charles Giattino, Duke University
Thomas K. Gilbert, UC Berkeley
Naman Goel, Artificial Intelligence Lab, EPFL
Sharad Goel, Stanford University
Judy Goldsmith, University of Kentucky
Jessica Goldthwaite, The Legal Aid Society

Naveen Sundar Govindarajulu, Rensselaer Polytechnic
Logan Graham, University of Oxford
Bradley Gram-Hansen, University of Oxford
Krishna P. Gummadi, Max Planck Institute
Andras Gyorgy, DeepMind
Gillian Hadfield, University of Toronto
Dylan Hadfield-Menell, UC Berkeley
Nika Haghtalab, Microsoft Research
The Anh Han,Teesside University
Tamir Hazan, Technion
Neil Heffernan IV, Shrewsbury High School
Patrick Helber, DFKI/TU Kaiserlautern
Jose Hernandez-Orallo, Univ. Politécnica de València
Michael Hind, MIT
Ayanna Howard, Georgia Institute of Technology
Wenjie Hu, Stanford University
Karen Huang, Harvard University
Sandy Huang, UC Berkeley
Clinton Hughes, The Legal Aid Society
Mark Ibrahim, Center for ML, Capital One
Ryan Blake Jackson, Colorado School of Mines
Sophie Jentzsch, TU Darmstadt
Ray Jiang, DeepMind
Adam Tauman Kalai, Microsoft Research
Ece Kamar, Microsoft Research
Subbarao Kambhampati, Arizona State University
Kristian Kersting, TU Darmstadt
Michael P. Kim, Stanford University
Mathhew Klenk, Palo Alto Research Center
Pushmeet Kohli, DeepMind
Veronika Kopackova, Czech Geological Survey
Dan Krane, Wright State University
Pierre Kreitmann, Google
Vigneshram Krishnamoorthy, Carnegie Mellon University
Lydia Kwong, Duke University
Himabindu Lakkaraju, Harvard University
Tor Lattimore, DeepMind
Tom Lenaerts, Université Libre de Bruxelles
Adam Lerer, Facebook
Jure Leskovec, Stanford University
Cyril Leung, University of British Columbia
Michael Lewis, University of Pittsburgh
Huao Li, University of Pittsburgh
Beishui Liao, Zhejiang University
Mark DM Leiserson, University of Maryland
Daniel Lim, Duke Kunshan University
Nicole Limtiaco, Google AI
Zhiyuan Lin, Stanford University
Yang Liu, UC Santa Cruz
David Lobell, Stanford University

Stephen Lorenz, Clarkson University
Melissa Louie, Center for ML, Capital One
Christine Luu, Google
Bertram Malle, Brown University
Ben Marafino, Stanford University
Nicholas Mattei, Tulane University
Jeanna Matthews, Clarkson University
Abigail Matthews, Clarkson University
Micki McGee, Fordham University
Daniel McNamara, Australian National Univ./CSIRO Data61
Kuldeep S. Meel, National University of Singapore
Wagner Meira Jr., Univ. Federal de Minas Gerais
Chunyan Miao, Nanyang Technological University
Stephanie Milani, University of Maryland
James Minton, Kansas State University
Yonatan Mintz, Georgia Institute of Technology
Ceena Modarres, Center for ML, Capital One
Shiwali Mohan, Palo Alto Research Center
Aleksandra Mojsilovic, MIT
Andrew Morgan, Cornell University
Arslan Munir, Kansas State University
Zairah Mustahsan, IBM Watson AI
Antonio Nicolo, University of Padova
Mariama Njie, Iona College
Mehrdad Nojoumian, Florida Atlantic University
Alejandro Noriega-Campero, MIT
Paul Novosad, Dartmouth College
Rune Nyrup, Leverhulme Centre for the FOI
Luca Oneto, University of Genoa
Cheng Soon Ong, Australian National Univ/CSIRO Data61
Michael Osborne, University of Oxford
John Paisley, Columbia University
Latha Palaniappan, Stanford University
Ravi Pandya, UC Berkeley
Corey Park, Florida Atlantic University
Jack Parker, Carnegie Mellon University
David Parkes, Harvard University
Rafael Pass, Cornell University
Jay Harshadbhai Patel, Stanford University
Tracy Pearl, Texas Tech University
Alex Pentland, MIT
Luiz Moniz Pereira, Universidade NOVA de Lisboa
Vincent Perot, Google AI
Alexander Peysakhovich, Facebook
Stephen Pfohl, Stanford University
Massimiliano Pontil, IIT
Hai Qian, Google
Goran Radanovic, Harvard University
Chad Rafetto, Duke University

Inioluwa Deborah Raji, University of Toronto
Hesham Rakha, Virginia Tech Transportation Institute
Karthikeyan Natesan Ramamurthy, IBM Research
Zoe-Alanah Robert, Stanford University
Fatima Rodriguez, Stanford University
Francesca Rossi, IBM
Constantin Rothkopf, TU Darmstadt
Daniela Rus, MIT
Nripsuta Saxena, University of Southern California
Matthias Scheutz, Tufts University
Patrick Schramowski, TU Darmstadt
Wilko Schwarting, MIT
Nigam Shah, Stanford University
Shervin Shahrdar, Florida Atlantic University
Adish Singla, Max Planck Inst. for Software Systems
Marija Slavkovik, University of Bergen
Ava Soleimany, Harvard University
Yi-Zhe Song, EECS-Queen Mary Univ. of London
Skyler Speakman, IBM Research
Daniel Susser, Pennsylvania State University
Nathaniel Swinger, Lexington High School
Katia Sycara, Carnegie Mellon University
Ankur Taly, Google AI
Zhongyi Tang, Stanford Inst. for Econ. Policy Research
Stefano Teso, Katholieke Universiteit Leuven
Wan-Hsiu Tsai, University of Miami
Shannon Vallor, Santa Clara University/Google
Leendert van der Torre, University of Luxembourg
Indhu Varatharajan, DLR Inst.for Planetary Research
Kush Varshney, IBM Research
Ellen Vitercik, Carnegie Mellon University
Karina Vold, Leverhulme Centre for the FOI
Dennis Wei, IBM Research
Sean Welsh, University of Canterbury
Ruchen Wen, Colorado School of Mines
Jess Whittlestone, Leverhulme Centre for the FOI
Tom Williams, Colorado School of Mines
Robert Williamson, Australian National Univ/CSIRO Data61
Lior Wolf, Facebook AI Research, Tel-Aviv University
Alison Woodruff, Google
Ava Thomas Wright, University of Georgia
Frances Yan, Palo Alto Research Center
Kumar Yogeeswaran, University of Canterbury
Han Yu, Nanyang Technological University
Muhammad Bilal Zafar, Max Planck Institute
Mohammed J. Zaki, Rensselaer Polytechnic Institute
Yongqing Zheng, Shandong University
James Zou, Stanford University

AIES-19 Conference Overview

On Saturday Jan. 26 at 7:30pm, we will have a **panel on "Perspectives on Responsible AI"** with Peter Hershock (moderator), Yolanda Gil, Huw Price, Francesca Rossi, and Dekai Wu. Location: *Spalding Building, Room 155, University of Hawaii - Manoa.* **The main program** (Jan. 27-28) takes place at the *Mid Pacific Conference Center at the Hilton Hawaiian Village, Coral 4.*

	Sunday, January 27	**Monday, January 28**
8:40 - 8:50	Opening remarks	
8:50 - 9:40	Invited talk **Ryan Calo (**Univ. of Washington)	Invited talk **Anca Dragan** (UC Berkeley)
9:40 - 10:00	Spotlight 1: *Normative Perspectives*	Spotlight 2: *Empirical Perspectives*
10:00 - 10:30	Coffee + Posters	Coffee + Posters
10:30 - 11:30	Session 1: *Algorithmic Fairness*	Session 6: *Social Science Models*
11:30 - 12:30	Session 2: *Norms and Explanations*	Session 7: *Measurement & Justice*
12:30 - 2:00	Lunch	Lunch
2:00 - 3:00	Session 3: *Artificial Agency*	Session 8: *AI for Social Good*
3:00 - 4:00	Session 4: *Autonomy & Lethality*	Session 9: *Human-Machine Interaction*
4:00 - 4:30	Coffee + Posters	Coffee + Posters
4:30 - 5:30	Session 5: *Rights & Principles* Spotlight 2: *Fairness+Explanations*	Invited talk **David Danks (**Carnegie Mellon) Closing remarks
5:40 - 6:30	Invited talk **Susan Athey** (Stanford University)	
7:00 - 9:00	Reception + Posters	