AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

The First AAAI/ACM Conference on AI, Ethics, and Society

February 1 – 3, 2018

Hilton New Orleans Riverside
New Orleans, Louisiana, 70130 USA

Program Guide

Organized by
AAAI, ACM, and ACM SIGAI

Sponsored by
Berkeley Existential Risk Initiative, DeepMind Ethics & Society, Future of Life Institute, IBM Research AI, Pricewaterhouse Coopers, Tulane University

AIES 2018 Conference Overview

	Thursday, February 1st	**Friday, February 2nd**	**Saturday, February 3rd**
	Tulane University	*Hilton Riverside*	*Hilton Riverside*
8:30-9:00		Opening	
9:00-10:00		Invited talk, AI: **Iyad Rahwan and Edmond Awad**, MIT	Invited talk, AI and jobs: **Richard Freeman**, Harvard
10:00-10:15		Coffee Break	Coffee Break
10:15-11:15		AI session 1	AI session 3
11:15-12:15		AI session 2	AI session 4
12:15-2:00		Lunch Break	Lunch Break
2:00-3:00		AI and law session 1	AI and philosophy session 1
3:00-4:00		AI and law session 2	AI and philosophy session 2
4:00-4:30		Coffee break	Coffee Break
4:30-5:30		Invited talk, AI and law: **Carol Rose**, ACLU	Invited talk, AI and philosophy: **Patrick Lin**, Cal Poly
5:30-6:30	6:00 – Panel 1: **What will Artificial Intelligence bring?**	Panel 2: **Prioritizing Ethical Considerations in AI: Who Sets the Standards?**	Invited talk: **The Venerable Tenzin Priyadarshi**, MIT
7:00	Reception	Conference reception	Closing

Acknowledgments

AAAI and ACM acknowledges and thanks the following individuals for their generous contributions of time and energy in the successful creation and planning of AIES 2018:

Program Chairs:
AI and jobs: Jason Furman (Harvard University)
AI and law: Gary Marchant (Arizona State University)
AI and philosophy: Huw Price (Cambridge University)
AI: Francesca Rossi (IBM and University of Padova)
Student Program Chair: Nicholas Mattei (IBM)
Website chair: Francisco Cruz (IIIA)
Public event chair: Brent Venable (Tulane University)
Standards' panel chair: John Havens (IEEE)
Local organization:
Carol Hamilton (Executive Director, AAAI)
Monique Abed (Conference Coordinator, AAAI)
AIES Steering Committee:
Vincent Conitzer (Duke University)
Subbarao Kambhampati (Arizona State University, AAAI)
Sven Koenig (University of Southern California, ACM SIGAI)
Francesca Rossi (IBM and University of Padova)
Bobby Schnabel (University of Colorado Boulder)

Program Committee:

Anna Alexandrova
Shahar Avin
Haris Aziz
Paula Boddington
Cameron Buckner
Stephen Cave
Vincent Conitzer
David Danks
Deven Desai
John Dickerson
Virginia Dignum
Kenny Easwaran
Judy Goldsmith
Dov Greenbaum
Joshua Greene
Maura R. Grossman
Thomas Icard
Hong Jiang
Stephen John
Irene Kitsara
Sven Koenig
Victoria Krakovna
Benjamin Kuipers
Ben Levinstein
Matthew Liao
Patrick Lin
Yang Liu
Gary Marcus
Nicholas Mattei
AJung Moon
Vincent Mueller
Iyad Rahwan
Stuart Russell
Bart Selman
Susan Schneider
Francis X. Shen
Biplav Srivastava
Milind Tambe
Shannon Vallor
Brent Venable
Wendell Wallach
Toby Walsh
Brian Williams
Andrew Keane Woods

Student program committee:

Emanuelle Burton
Maria Chang
Deven Desai
John Dickerson
Judy Goldsmith
Benjamin Kuipers
Andrea Loreggia
Michael Rovatsos
Kartik Talamadupula
K. Brent Venable

Additional Reviewers:

Kartik Talamadupula
Aaron Schlenker
Aida Rahmattalabi
Andrea Loreggia
Andreas Hofmann
Andreas Hofmann
Anja Dahlmann
Ashkan Jasour
Avinash Balakrishnan
Barton Lee
Cory Siler
Cristina Cornelio
Dominik Peters
Edward J. Lee
François Garillot
Hang Ma
Hong Xu
Jana Schaich Borg
Jason Millar
Jiaoyang Li
Kamran Najeebullah
Kristel Clayville
Kshitij Fadnis
Liron Cohen
Lok Chan
Markus Brill
Max Kramer
Nikhil Bhargava
Omer Lev
Rene Weber
Sara Magliacane
Shahrzad Gholami
Shalaleh Rismani
Shawn Schaffert
Shokoofeh Pourmehr
Tansel Uras
Yuening Zhang

Registration

Onsite registration and badge pick-up will be located in the foyer of The District, third floor of the Hilton New Orleans Riverside Hotel, Two Poydras Street, New Orleans, Louisiana, 70130, USA. Registration hours will be Friday, Saturday, February 2–3, 7:30 AM – 5:00 PM. All attendees must pick up their registration packets/badges for admittance to programs.

Wifi

To access the wifi in the conference rooms, please use the following login information:
Network: Hilton Meetings, Password: DisneyAAAI18

Social events

Public reception, offered by Tulane University
Feb. 1st, 6:00pm-8:30pm
Freeman Auditorium, Woldenberg Art Center
7018-7098 Plum St, New Orleans, LA 70118
Open to all conference attendants and the public

Student lunch
Feb.2nd, 12:15pm-2:00pm
Mulate's, 201 Julia Street, New Orleans, LA
For accepted students, program chairs, panel chairs, student program chair, invited speakers, sponsors

Conference reception, offered by DeepMind Ethics & Society
Feb.2nd, 7:00pm – 9pm
St. James Ballroom, 3rd floor, Hilton New Orleans Riverside
Open to all conference attendants

AIES 2018 Technical Program

February 1st

Tulane University

6:00 - 8:30 PM

Panel 1: What will Artificial Intelligence bring? Discussing the advent and consequences of superhuman intelligence

Moderator: Brent Venable (Tulane University)

Panelists:

- Paula Boddington (Oxford University)
- Wendell Wallach (Yale University)
- Jason Furman (Harvard University)
- Peter Stone (UT Austin)

As AI is becoming more pervasive in our life, its impact on society is more significant and concerns and issues are raised regarding aspects such as value alignment, data handling and bias, regulations, and workforce displacement. Recognizing the importance of providing scientifically sound and reliable information on this topic, the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society will open its program with a panel open to the public. World class researchers from different disciplines and best selling authors will elaborate on the impact of AI on modern society and will answer questions from the public.

Tulane University
Freeman Auditorium, Woldenberg Art Center
7018-7098 Plum St, New Orleans, LA 70118
Open to public
Followed by reception, offered by Tulane University

February 2nd

Quarterdeck Room, Riverside Complex,
Hilton New Orleans Riverside

8:30 - 9:00 AM

Opening and best paper award

9:00 -10:00 AM

Invited talk, AI:
The Moral Machine Experiment: 40 Million Decisions and the Path to Universal Machine Ethics

Iyad Rahwan and Edmond Awad (MIT)
Chair: Francesca Rossi

We describe the Moral Machine, an internet-based serious game exploring the

many-dimensional ethical dilemmas faced by autonomous vehicles. The game enabled us to gather 40 million decisions from 3 million people in 200 countries/territories. We report the various preferences estimated from this data, and document interpersonal differences in the strength of these preferences. We also report cross-cultural ethical variation and uncover major clusters of countries exhibiting substantial differences along key moral preferences. These differences correlate with modern institutions, but also with deep cultural traits. We discuss how these three layers of preferences can help progress toward global, harmonious, and socially acceptable principles for machine ethics.

10:00 - 10:15 AM
Coffee Break

10:15 - 11:15 AM
AI session 1 (value alignment)
Chair: Jason Furman

- Maite Lopez-Sanchez, Marc Serramia, Juan Antonio Rodriguez-Aguilar, Carlos Ansotegui, Javier Morales and Michael Wooldridge. Exploiting moral values to choose the right norms
- Andrea Loreggia, Nicholas Mattei, Francesca Rossi and Kristen Brent Venable. Preferences and Ethical Principles in Decision Making
- Ryan Carey. Incorrigibility in the CIRL Framework
- Collin Johnson and Benjamin Kuipers. Socially-Aware Navigation Using Topological Maps and Social Norm Learning

11:15 AM - 12:15 PM
AI session 2 (bias and fairness)
Chair: Jason Furman

- John Li, Lucas Dixon, Nithum Thain, Lucy Vasserman and Jeffrey Sorensen. Measuring and Mitigating Unintended Bias in Text Classification
- Naman Goel, Mohammad Yaghini and Boi Faltings. Non-Discriminatory Machine Learning through Convex Fairness Criteria
- Edward Raff, Jared Sylvester and Steven Mills. Fair Forests: Regularized Tree Induction to Minimize Model Bias
- Sarah Tan, Rich Caruana, Giles Hooker and Yin Lou. Detecting Bias in Black-Box Models Using Transparent Model Distillation

12:15 - 2:00 PM
Lunch Break

2:00 - 3:00 PM
AI and law session 1 (Reponsibility)
Chair: Gary Marchant

- Cindy Grimm, Bill Smart and Woodrow Hartzog. Using Education as a Model to Capture Good-Faith Effort for Autonomous Systems
- Daniel Tobey. Software Malpractice in the Age of AI: A Guide for the Wary Tech Company
- Lav Varshney and Deepak Somaya. Embodiment, Anthropomorphism, and Intellectual Property Rights for AI Creations
- Sjur Kristoffer Dyrkolbotn, Truls Pedersen and Marija Slavkovik. On the distinction between implicit and explicit ethical agency

3:00 - 4:00 PM
AI and law session 2 (Governance)
Chair: Gary Marchant

- Olivia Johanna Erdelyi and Judy Goldsmith. Regulating Artificial Intelligence: Proposal for a Global Solution
- Wendell Wallach and Gary Marchant. An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics
- Matthijs Maas. Regulating for 'normal AI accidents': operational lessons for the responsible governance of AI deployment
- Stephen Cave and Sean O Heigeartaigh. An AI Race: Rhetoric and Risks

4:00 - 4:30 PM
Coffee Break

4:30 - 5:30 PM
Invited talk, AI and law:
AI, Civil Rights, and Civil Liberties: Can Law Keep Pace with Technology?
Carol Rose (ACLU)
Chair: Gary Marchant

At the dawn of this era of human-machine interaction, humans beings have an opportunity to shape fundamentally the ways in which machine learning will expand or contract the human experience, both individually and collectively. As efforts to develop guiding ethical principles and legal constructs for human-machine interaction move forward, how do we address not only what we do with AI, but also the question of who gets to decide and how? Are guiding principles of Liberty and Justice for All still relevant? Does a new era require new models of open leadership and collaboration around law, ethics, and AI?

5:30 - 6:30 PM
Panel 2: Prioritizing Ethical Considerations in Intelligent and Autonomous Systems - Who Sets the Standards?
Moderator: John Havens (IEEE)
Panelists*:*

- Takashi Egawa (NEC Corporation)

- Simson L. Garfinkel (USACM)
- John C. Havens (IEEE)
- Dan Palmer (The British Standards Institution)
- Annette Reilly (IEEE)

Traditionally, product engineering standards have focused squarely on issues of technological interoperability and safety. While considerations along these lines can be extremely complex, standards have not overtly tackled applied ethical issues directly when being created. Today, however, organizations like The British Standards Institute (BSI) with their BS 8611: Robots and robotic devices: Guide to the ethical design and application of robots and robotic systems Standard, and IEEE's P7000 suite of thirteen standardization projects are creating a new paradigm in the field. While dealing with intelligent and autonomous technologies, these standards and standardization projects are also providing detailed guidelines or requirements to help organizations institute new levels of transparency, accountability and traceability that can build trust and maximize innovation while avoiding negative unintended consequences.

7:00 PM
Conference reception, offered by DeepMind Ethics & Society

February 3rd

Quarterdeck Room, Riverside Complex,
Hilton New Orleans Riverside

9:00 – 10:00 AM
Invited talk, AI and jobs:
The Great AI/Robot Jobs Scare: reality of automation fear redux
Richard Freeman (Harvard University)
Chair and discussant: Jason Furman

This talk will consider the impact of AI/robots on employment, wages and the future of work more broadly. We argue that we should focus on policies that make AI robotics technology broadly inclusive both in terms of consumption and ownership so that billions of people can benefit from higher productivity and get on the path to the coming age of intolerable abundance.

10:00 - 10:15 AM
Coffee break

10:15 - 11:15 AM
AI session 3 (ethical issues and models)
Chair: Francesca Rossi

- Bobbie Eicher, Lalith Polepeddi and Ashok Goel. Jill Watson Doesn't Care if You're Pregnant: Grounding AI Ethics in Empirical Studies

- Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe and Joelle Pineau. Ethical Challenges in Data-Driven Dialogue Systems
- Dieter Vanderelst and Alan Winfield. The Dark side of Ethical Robots
- Richard Kim, Max Kleiman-Weiner, Andres Abeliuk, Edmond Awad, Sohan Dsouza, Josh Tenenbaum and Iyad Rahwan. A Computational Model of Commonsense Moral Decision Making

11:15 AM - 12:15 PM
AI session 4 (transparency and social good)
Chair: Francesca Rossi

- Fan-Yun Sun, Yen-Yu Chang, Yueh-Hua Wu and Shou-De Lin. Designing Non-greedy Reinforcement Learning Agents with Diminishing Reward Shaping
- Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar and Katia Sycara. Transparency and Explanation in Deep Reinforcement Learning Neural Networks
- Sungyong Seo, Hau Chan, P. Jeffrey Brantingham, Jorja Leap, Phebe Vayanos, Milind Tambe and Yan Liu. Partially Generative Neural Networks for Gang Crime Classification with Partial Information
- Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Fabricio Benevenuto, Krishna P. Gummadi and Adrian Weller. Purple Feed: Identifying High Consensus News Posts on Social Media

12:15 - 2:00 PM
Lunch Break

2:00 - 3:00 PM
AI and philosophy session 1 (Ethics and policy)
Chair: Huw Price

- Mathieu D'Aquin, Pinelopi Troullinou, Noel O'Connor, Aindrias Cullen, Gráinne Faller and Louise Holden. Towards an “Ethics by Design” methodology for AI research projects
- Ross Gruetzemacher. Rethinking AI Strategy and Policy as Entangled Super Wicked Problems
- Alex John London and David Danks. Regulating Autonomous Vehicles: A Policy Proposal
- Max Kramer, Jana Schaich Borg, Vincent Conitzer and Walter Sinnott-Armstrong. When Do People Want AI to Make Decisions?

3:00 - 4:00 PM
AI and philosophy session 2 (Machine Ethics)

Chair: Huw Price

- Daniel Estrada. Value Alignment, Fair Play, and the Rights of Service Robots
- Michael Scheessele. A framework for grounding the moral status of intelligent machines
- Emily Larosa and David Danks. Impacts on Trust of Healthcare AI
- John Hooker and Tae Wan Kim. Toward Non-Intuition-Based Machine Ethics

4:00 - 4:30 PM
Coffee Break

4:30 - 5:30 PM
Invited talk, AI and philosophy:
AI Decisions, Risk, and Ethics: Beyond Value Alignment
Patrick Lin (California Polytechnic State University)
Chair: Huw Price

When we think about the values AI should have in order to make right decisions and avoid wrong ones, there's a large but hidden *third* category to consider: decisions that are not-wrong but also not-right. This is the grey space of judgment calls, and just having good values might not help as much as you'd think here. I'll use autonomous cars as my case study here, with lessons for broader AI: ethical dilemmas can arise in everyday scenarios such as lane positioning and navigation, not just in crazy crash scenarios. This is the space where one good value might conflict with another good value, and there's no "right" answer or even broad consensus on an answer; so it's important to recognize the hard cases—which are potential limits—in the study of AI ethics.

5:30 – 6:30 PM
Invited talk:
Ethics, Empathy and Extended Intelligence
The Venerable Tenzin Priyadarshi (MIT)
Chair: Huw Price

As advent of AI mirrors and augments the desire for exponential growth and global domination, are we asking the right set of questions to facilitate development of a wholesome ethical framework to guide the development and deployment of AI systems? What ought to be the role of humans? Are our social contracts and policy frameworks in place to usher a new wave of AI related developments? Are we ready to adapt? Can defining AI via a non-reductionist and distributed phenomenological lens or Extended Intelligence help?

6:30 PM
Closing

Student Program

Out of 65 student applications, we selected 20 students to participate in the conference. These students received a free registration and got a travel grant. Moreover, the accepted students will

- present a poster (during the student poster session, in the afternoon of Feb.2nd),
- meet senior members of the community and representatives of the sponsoring companies,
- be invited at the student lunch on Feb.2nd
- have the opportunity to publish a 2-page extended abstract of their work in the proceedings

Sessions and presentation length

Each session is 1 hour long and includes 4 presentations of 10' each (including questions), followed by a 20 minutes panel with the authors of the presented papers as panelists. The session chair will moderate the panel.

Poster sessions

Quarterdeck Room (3rd section), Riverside Complex,
Hilton New Orleans Riverside

Poster boards are 30" x 40" (vertical) foamcore, mounted on easels. AAAI will provide push pins for authors to mount their posters.

There will be 4 poster sessions, with 20 posters each:

Poster session 1: morning of Feb.2nd

- John Li, Lucas Dixon, Nithum Thain, Lucy Vasserman and Jeffrey Sorensen. Measuring and Mitigating Unintended Bias in Text Classification
- Daniel Estrada. Value Alignment, Fair Play, and the Rights of Service Robots
- Olivia Johanna Erdelyi and Judy Goldsmith. Regulating Artificial Intelligence: Proposal for a Global Solution
- Cindy Grimm, Bill Smart and Woodrow Hartzog. Using Education as a Model to Capture Good-Faith Effort for Autonomous Systems
- Maite Lopez-Sanchez, Marc Serramia, Juan Antonio Rodriguez-Aguilar, Carlos Ansotegui, Javier Morales and Michael Wooldridge. Exploiting moral values to choose the right norms
- Daniel Tobey. Software Malpractice in the Age of AI: A Guide for the Wary

Tech Company

- Naman Goel, Mohammad Yaghini and Boi Faltings. Non-Discriminatory Machine Learning through Convex Fairness Criteria
- Edward Raff, Jared Sylvester and Steven Mills. Fair Forests: Regularized Tree Induction to Minimize Model Bias
- Michael Scheessele. A framework for grounding the moral status of intelligent machines
- Ross Gruetzemacher. Rethinking AI Strategy and Policy as Entangled Super Wicked Problems
- Upol Ehsan, Brent Harrison, Larry Chan and Mark Riedl. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations
- Ryan Carey. Incorrigibility in the CIRL Framework
- Max Kramer, Jana Schaich Borg, Vincent Conitzer and Walter Sinnott-Armstrong. When Do People Want AI to Make Decisions?
- John Hooker and Tae Wan Kim. Toward Non-Intuition-Based Machine Ethics
- Lav Varshney and Deepak Somaya. Embodiment, Anthropomorphism, and Intellectual Property Rights for AI Creations
- Sungyong Seo, Hau Chan, P. Jeffrey Brantingham, Jorja Leap, Phebe Vayanos, Milind Tambe and Yan Liu. Partially Generative Neural Networks for Gang Crime Classification with Partial Information
- Sarah Tan, Rich Caruana, Giles Hooker and Yin Lou. Detecting Bias in Black-Box Models Using Transparent Model Distillation
- Fan-Yun Sun, Yen-Yu Chang, Yueh-Hua Wu and Shou-De Lin. Designing Non-greedy Reinforcement Learning Agents with Diminishing Reward Shaping
- Dieter Vanderelst and Alan Winfield. The Dark side of Ethical Robots
- Bobbie Eicher, Lalith Polepeddi and Ashok Goel. Jill Watson Doesn't Care if You're Pregnant: Grounding AI Ethics in Empirical Studies

Poster session 2 (student posters): afternoon of Feb.2nd

- Stefano Tedeschi. Accountable Agents and Where to Find Them
- Eva Thelisson. Toward a computational sustainability for AI/ML to foster Responsibility
- Ilse Verdiesen. The Design of Human Oversight in Autonomous Weapons
- Jin Xu. Overtrust of Robots in High-Risk Scenarios
- Barton Lee. A win for society! Conquering barriers to fair elections
- Martin Strobel. An axiomatic approach to explain Computer generated decisions
- Priel Levy. Optimal Contest Design for Multi-Agent Systems
- Gong Chen. Nurturing the Companion ChatBot
- Daniel Kasenberg. Learning and obeying conflicting norms in stochastic domains

- Marc Serramia. Ethics in norm decision making
- Joseph Blass. Legal, Ethical, Customizable Artificial Intelligence
- Rediet Abebe. Computational Perspectives on Social Good and Access to Opportunity
- Gagan Bansal. Explanatory Dialogs: Towards Actionable, Interactive Explanations
- Shiva Kaul. Speed and accuracy are not enough! Trustworthy machine learning
- Bobbie Eicher. Giving AI a Theory of Mind
- Fulton Wang. Fulton Wang's Application to the AIES Student Program
- Cassandra Carley. Balancing Privacy and Utility with Pattern Based Activity Detection
- Colin Garvey. AI Risk Mitigation through Democratic Governance
- Sarah Tan. Interpretable Approaches to Detect Bias in Black-Box Models
- Lily Hu. Justice Beyond Utility in Artificial Intelligence

Poster session 3: morning of Feb.3nd

- Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Fabricio Benevenuto, Krishna P. Gummadi and Adrian Weller. Purple Feed: Identifying High Consensus News Posts on Social Media
- Alex John London and David Danks. Regulating Autonomous Vehicles: A Policy Proposal
- Mathieu D'Aquin, Pinelopi Troullinou, Noel O'Connor, Aindrias Cullen, Gráinne Faller and Louise Holden. Towards an "Ethics by Design" methodology for AI research projects
- Matthijs Maas. Regulating for 'normal AI accidents': operational lessons for the responsible governance of AI deployment
- Sjur Kristoffer Dyrkolbotn, Truls Pedersen and Marija Slavkovik. On the distinction between implicit and explicit ethical agency
- Richard Kim, Max Kleiman-Weiner, Andres Abeliuk, Edmond Awad, Sohan Dsouza, Josh Tenenbaum and Iyad Rahwan. A Computational Model of Commonsense Moral Decision Making
- Collin Johnson and Benjamin Kuipers. Socially-Aware Navigation Using Topological Maps and Social Norm Learning
- Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar and Katia Sycara. Transparency and Explanation in Deep Reinforcement Learning Neural Networks
- Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe and Joelle Pineau. Ethical Challenges in Data-Driven Dialogue Systems
- Stephen Cave and Sean O Heigeartaigh. An AI Race: Rhetoric and Risks
- Andrea Loreggia, Nicholas Mattei, Francesca Rossi and Kristen Brent Venable. Preferences and Ethical Principles in Decision Making
- Alan Wagner. An Autonomous Architecture that Protects the Right to Privacy

- Hau Chan, Long Tran-Thanh, Bryan Wilder, Eric Rice, Phebe Vayanos and Milind Tambe. Utilizing Housing Resources for Homeless Youth Through the Lens of Multiple Multi-Dimensional Knapsacks
- Habib Karbasian, Hemant Purohit, Rajat Handa, Aqdas Malik and Aditya Johri. Real-Time Inference of User Types to Assist with more Inclusive and Diverse Social Media Activism Campaigns
- Sandeep Konam, Ian Quah, Stephanie Rosenthal and Manuela Veloso. Understanding Convolutional Networks with APPLE: Automatic Patch Pattern Labeling for Explanation
- Piercosma Bisconti Lucidi and Daniele Nardi. Companion Robots: the Hallucinatory Danger of Human-Robot Interactions
- Martijn van Otterlo. From Algorithmic Black Boxes to Adaptive White Boxes: Declarative Decision-Theoretic Ethical Programs as Codes of Ethics
- Jianxin Zhao, Richard Mortier, Jon Crowcroft and Liang Wang. Privacy-preserving Machine Learning Based Data Analytics on Edge Devices
- Daniel Kasenberg and Matthias Scheutz. Inverse norm conflict resolution
- Biplav Srivastava and Francesca Rossi. Towards Composable Bias Rating of AI Systems

Poster session 4: afternoon of Feb.3nd

- Golnoosh Farnadi, Behrouz Babaki and Lise Getoor. Fairness in Relational Domains
- Amit Chopra and Munindar Singh. Sociotechnical Systems and Ethics in the Large
- Shiva Kaul. Margins and opportunity
- Shivaram Kalyanakrishnan, Rahul Panicker, Sarayu Natarajan and Shreya Rao. Opportunities and Challenges for Artificial Intelligence in India
- Brian Zhang, Blake Lemoine and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning
- Nicholas Mattei, Abdallah Saffidine and Toby Walsh. Fairness in Deceased Organ Matching
- Lydia Manikonda, Aditya Deotale and Subbarao Kambhampati. What's up with Privacy?: User Preferences and Privacy Concerns in Intelligent Personal Assistants
- Yuanshuo Zhao, Ioana Baldini, Prasanna Sattigeri, Inkit Padhi, Yoong Keok Lee and Ethan Smith. Data Driven Platform for Organizing Scientific Articles Relevant to Biomimicry
- Nolan P. Shaw, Andreas Stöckel, Ryan W. Orr, Thomas F. Lidbetter and Robin Cohen. Towards Provably Moral AI Agents in Bottom-up Learning Frameworks
- Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel and Aaron Roth. Meritocratic Fairness for Infinite and Contextual Bandits
- Eduardo Ivan Velazquez Richards, Mehrdad Yazdani and Pablo Suárez-Serrato. Socialbots supporting human rights
- Virginia Dignum, Matteo Baldoni, Cristina Baroglio, Maurizio Caon, Raja

Chatila, Louise A. Dennis, Gonzalo Génova, Malte Kliess, Maite Lopez-Sanchez, Roberto Micalizio, Juan Pavon, Marija Slavkovik, Matthijs Smakman, Marlies van Steenbergen, Stefano Tedeschi, Leon van der Torre, Serena Villata and Tristan de Wildt. Ethics by Design: Necessity or Curse?

- Kyle Hundman, Thamme Gowda, Mayank Kejriwal and Benedikt Boecking. Always Lurking: Understanding and Mitigating Bias in Online Human Trafficking Detection
- Marisa Vasconcelos, Bernardo Goncalves and Carlos Henrique Cardonha. Modeling Epistemological Principles for Bias Mitigation in AI Systems: An Illustration in Hiring Decisions
- Emily Larosa and David Danks. Impacts on Trust of Healthcare AI
- Haris Aziz and Barton Lee. Sub-committee Approval Voting and Generalized Justified Representation Axioms
- Wendell Wallach and Gary Marchant. An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics
- Daniel Kasenberg, Thomas Arnold and Matthias Scheutz. Norms, Rewards, and the Intentional Stance: Comparing Machine Learning Approaches to Ethical Training
- Edvard Pires Bjørgen, Simen Øvervatn Madsen, Therese Skaar Bjørknes, Fredrik Vonheim Heimsæter, Robin Håvik, Morten Linderud, Per-Niklas Longberg, Louise Dennis and Marija Slavkovik. Cake, death, and trolleys: dilemmas as benchmarks of ethical decision-making
- Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P. Dickerson and Vincent Conitzer. Adapting a Kidney Exchange Algorithm to Align with Human Values

Best paper award (sponsored by the Partnership on AI)

The AIES 2018 best paper award is shared between the following two papers:

Transparency and Explanation in Deep Reinforcement Learning Neural Networks, Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar, Katia Sycara
For AI systems to be accepted and trusted, the users should be able to understand the reasoning process of the system and to form coherent explanations of the systems decisions and actions. This paper presents a novel and general method to provide a visualisation of internal states of deep reinforcement learning models, thus enabling the formation of explanations that are intelligible to humans.

An AI Race: Rhetoric and Risks, Stephen Cave, Seán S ÓhÉigeartaigh
The rhetoric of the race for strategic advantage is increasingly being used with regard to the development of AI. This paper assesses the potential risks of the AI race narrative, explores the role of the research community in responding to

these risks, and discusses alternative ways to develop AI in a collaborative and responsible way.

Speakers

Edmond Awad (MIT), Invited speaker, Feb. 2nd, 9:00am
Edmond Awad is a Postdoctoral Associate at the Scalable Cooperation group led by Iyad Rahwan at MIT Media Lab. Born and raised in Syria, Edmond received his bachelor degree (2007) from Tishreen University (Syria) in Informatics Engineering. In 2009, he moved to UAE where at Masdar Institute, he completed a master's degree (2011) in Computing and Information Science with a research topic in Multi-agent Systems, before completing a PhD (2015) in Argumentation and Multi-agent systems. In 2015, Edmond joined the Scalable Cooperation group at MIT Media Lab as a graduate student and a research assistant. During his second master's degree, Edmond co-developed Moral Machine, a website that gather human decisions on moral dilemmas faced by driverless cars. The website has been visited by over 3 million users, who contributed their judgements on 40 million dilemmas. Edmond's work has been covered in major media outlets like The Times, LA Times, Der Spiegel (DE), and El Pais (ES). Edmond's research interest are in the areas of AI, Ethics, Computational Social Science and Multi-agent Systems.

Paula Boddington (Oxford University), Panelist for panel 1, Feb1st 6pm, Tulane University
Paula Boddington has been working in the Department of Computer Science at Oxford University on a project investigating the possibilities of developing codes of ethics for artificial intelligence, where she has been funded by the Future of Life Institute. She is a philosopher by background, and has worked extensively on questions in applied philosophy, including medical ethics and ethical questions in genetics and genomics. Her book "Towards a Code of Ethics for Artificial Intelligence" was published in December 2017.

Jason Furman (Harvard University), Panelist for panel 1, Feb1st 6pm, Tulane University
Jason Furman is Professor of the Practice of Economic Policy at Harvard Kennedy School (HKS). He is also nonresident senior fellow at the Peterson Institute for International Economics. This followed eight years as a top economic adviser to President Obama, including serving as the 28th Chairman of the Council of Economic Advisers from August 2013 to January 2017, acting as both President Obama's chief economist and a member of the cabinet. During this time Furman played a major role in most of the major economic policies of the Obama Administration. Previously Furman held a variety of posts in public policy and research. In public policy, Furman worked at both the Council of Economic Advisers and National Economic Council during the Clinton administration and also at the World Bank. In research, Furman was a Director of the Hamilton Project and Senior Fellow at the Brookings Institution and also has served in visiting positions at various universities, including NYU's Wagner Graduate

School of Public Policy. Furman has conducted research in a wide range of areas, including fiscal policy, tax policy, health economics, Social Security, technology policy, and domestic and international macroeconomics. In addition to articles in scholarly journals and periodicals, Furman is the editor of two books on economic policy. Furman holds a Ph.D. in economics from Harvard University.

Takashi Egawa (NEC Corporation), Panelist for panel 2, Feb. 2nd, 5:30pm
Takashi Egawa is an expert of standardization, in particular telecommunication technology and AI/S. He contributed to the standardization of Next Generation Network, disaster telecommunication, and Future Networks as the chair of Focus Group on Future Networks, the chair of Joint Coordination Activity of Software Defined Networks (SDN), the Rapporteur of SDN, and others in ITU-T. He is now working in NEC Corporation and is responsible for the standardization of AI/S, in particular Ethical, Legal, Social Issues (ELSI) of AI/S. In IEEE, he serves as the secretary of IEEE P7001 (Transparency of Autonomous Systems).

Richard Freeman (Harvard University), Invited speaker, Feb.3rd, 9:00am
Richard B. Freeman holds the Herbert Ascherman Chair in Economics at Harvard University. He is currently serving as Faculty co-Director of the Labor and Worklife Program at the Harvard Law School, and is Senior Research Fellow in Labour Markets at the London School of Economics' Centre for Economic Performance. He directs the National Bureau of Economic Research / Science Engineering Workforce Projects, and is Co-Director of the Harvard Center for Green Buildings and Cities.

Simson Garfinkel (ACM U.S. Public Policy Council), Panelist for panel 2, Feb.2nd, 5:30pm
Simson L. Garfinkel is a member of the Association for Computing Machinery's Public Policy Council and the co-chair of Council's Ad Hoc Working Group on Algorithms. His current research interests include privacy in big data, cybersecurity and usability. He holds seven US patents and has published dozens of research articles in computer security and digital forensics. He is an ACM Fellow and a member of the National Association of Science Writers.

John C. Havens (IEEE), Panelist for panel 2, Feb.2nd, 5:30pm
John C. Havens is Executive Director of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems that has two outputs – the creation and iteration of *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* and the identification and recommendation of ideas for Standards Projects focused on prioritizing ethical considerations in AI/AS. Currently there are thirteen approved Standards Working Groups in the IEEE P7000™ series.
John is also author of: *Heartificial Intelligence: Embracing Our Humanity To Maximize Machines* and *Hacking Happiness: Why Your Personal Data Counts and How Tracking it Can Change the World*.

Patrick Lin (California Polytechnic State University), Invited speaker, Feb. 3rd, 4:30pm

Patrick Lin, PhD, is the director of the Ethics + Emerging Sciences Group, based at California Polytechnic State University, San Luis Obispo, where he is a philosophy professor. Current affiliations include Stanford Law School, Notre Dame, World Economic Forum's Global Future Council on AI and Robotics, and the 100-Year Study on AI. Previous affiliations include Stanford Engineering, US Naval Academy, Dartmouth College, and UNIDIR. He is well published in technology ethics, with five books that include *Robot Ethics* (MIT Press, 2012) and *Robot Ethics 2.0* (Oxford University Press, 2017), as well as several funded policy reports on military robotics, cyberwarfare, and enhanced warfighters. Dr. Lin regularly gives invited briefings to industry, media, and governments worldwide; and he teaches courses in ethics, technology, and law.

The Venerable Tenzin Priyadarshi (MIT), Invited speaker, Feb.3rd, 5:30pm

The Venerable Tenzin Priyadarshi is an innovative thinker, philosopher, educator and a polymath monk. He is Director of the Ethics Initiative at the MIT Media Lab and President & CEO of The Dalai Lama Center for Ethics and Transformative Values at the Massachusetts Institute of Technology, a center dedicated to inquiry, dialogue, and education on the ethical and humane dimensions of life. The Center is a collaborative and nonpartisan think tank, and its programs emphasize responsibility and examine meaningfulness and moral purpose between individuals, organizations, and societies. Six Nobel Peace Laureates serve as The Center's founding members and its programs run in several countries and are expanding. Venerable Tenzin's unusual background encompasses entering a Buddhist monastery at the age of ten and receiving graduate education at Harvard University with degrees ranging from Philosophy to Physics to International Relations. He is a Tribeca Disruptive Fellow and a Fellow at the Center for Advanced Study in Behavioral Sciences at Stanford University. Venerable Tenzin serves on the boards of number of academic, humanitarian, and religious organizations. He is the recipient of several recognitions and awards, and received Harvard's Distinguished Alumni Honors for his visionary contributions to humanity.

Iyad Rahwan (MIT), Invited speaker, Feb. 2nd, 9:00am

Iyad Rahwan is the AT&T Career Development Professor and an Associate Professor of Media Arts & Sciences at the MIT Media Lab, where he leads the Scalable Cooperation group. A native of Aleppo, Syria, Rahwan holds a PhD from the University of Melbourne, Australia, and is an affiliate faculty at the MIT Institute of Data, Systems and Society (IDSS). Rahwan's work lies at the intersection of the computer and social sciences, with a focus on collective intelligence, large-scale cooperation, and the social aspects of Artificial Intelligence. He led the winning team in the US State Department's Tag Challenge, using social media to locate individuals in remote cities within 12 hours using only their mug shots. Recently, he crowdsourced 40 million decisions

from people worldwide about the ethics of AI systems. Rahwan's work appeared in major academic journals, including Science and PNAS, and features regularly in major media outlets, including the New York Times, The Economist, and the Wall Street Journal.

Annette Reilly (IEEE)**,** Panelist for panel 2, Feb.2nd, 5:30pm
Annette Reilly, Ph.D., IEEE Senior Life Member and IEEE Computer Society Golden Core, ACM Life Member, CSDP, CSEP-ACQ and INCOSE member, PMP, and STC Fellow, has led the development and harmonization of IEEE and ISO/IEC standards for systems and software engineering and information management for over 30 years. She retired from a 31-year career at Lockheed Martin, where she held a variety of responsibilities for proposal management, engineering management, systems engineering, information management, and technical documentation. Dr. Reilly received a B.A. from Rice University, M.A. and Ph.D. from Brandeis University, and an MIS from The George Washington University.

Carol Rose (ACLU), Invited speaker, Feb. 2nd, 4:30pm
Carol Rose is the Executive Director of the ACLU of Massachusetts (www.aclum.org), a nonpartisan organization with over 85,000 members and supporters in Massachusetts (alongside more than 1.6 million ACLU members nationwide) that integrates litigation, legislation, traditional and social media, and community-based movement-building to promote civil rights and defend civil liberties. A journalist and lawyer, Rose in 2013 launched the ACLU of Massachusetts' "Technology for Liberty & Justice for All" initiative, a $7-milllion program focused on the civil liberties implications and civil rights promise of new technology, developed in combination with a movement-building approach to law reform. Rose is a frequent speaker on technology and civil liberties issues, including the 2014 White House conference on big data privacy at MIT and the 2016 Forum on Data Privacy hosted by the Internet Policy Research Initiative at MIT. She is a member of the Board of Directors of the Partnership on Artificial Intelligence. She is a graduate of Stanford University (BSc 1983), the London School of Economics (MSc 1985), and Harvard Law School (JD 1996).

Peter Stone (UT Austin), Panelist for panel 1, Feb1st 6pm, Tulane University
Dr. Peter Stone is the David Bruton, Jr. Centennial Professor and Associate Chair of Computer Science, as well as Chair of the Robotics Portfolio Program, at the University of Texas at Austin. In 2013 he was awarded the University of Texas System Regents' Outstanding Teaching Award and in 2014 he was inducted into the UT Austin Academy of Distinguished Teachers, earning him the title of University Distinguished Teaching Professor. Professor Stone's research interests in Artificial Intelligence include machine learning (especially reinforcement learning), multiagent systems, robotics, and e-commerce. Professor Stone received his Ph.D in Computer Science in 1998 from Carnegie Mellon University. From 1999 to 2002 he was a Senior Technical Staff Member in the Artificial Intelligence Principles Research Department at AT&T Labs -

Research. He is an Alfred P. Sloan Research Fellow, Guggenheim Fellow, AAAI Fellow, IEEE Fellow, Fulbright Scholar, and 2004 ONR Young Investigator. In 2003, he won an NSF CAREER award for his proposed long term research on learning agents in dynamic, collaborative, and adversarial multiagent environments, in 2007 he received the prestigious IJCAI Computers and Thought Award, given biannually to the top AI researcher under the age of 35, and in 2016 he was awarded the ACM/SIGAI Autonomous Agents Research Award. Professor Stone co-founded Cogitai, Inc., a startup company focused on continual learning, in 2015, and currently serves as President and COO.

Wendell Wallach (Yale University), Panelist for panel 1, Feb1st 6pm, Tulane University

Wendell Wallach is senior advisor to The Hastings Center. He is also a scholar, consultant, and author at Yale University's Interdisciplinary Center for Bioethics, where he has chaired Technology and Ethics studies for the past eleven years. His latest book, a primer on emerging technologies, is entitled, A Dangerous Master: How to keep technology from slipping beyond our control. In addition, he co-authored (with Colin Allen) Moral Machines: Teaching Robots Right From Wrong. He received the World Technology Award for Ethics in 2014 and for Journalism and Media in 2015, as well as a Fulbright Research Chair at the University of Ottawa in 2015-2016. The World Economic Forum appointed Mr. Wallach co-chair of its Global Future Council on Technology, Values, and Policy for the 2016-2018 term.

Organizing Associations

AIES 2018 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 awardsupport 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

AIES 2018 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.

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.

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.

IBM Research AI

http://www.research.ibm.com

At IBM Research, we invent things that matter to the world. Today, we are pioneering promising and disruptive technologies that will transform industries and society, including the future of AI, blockchain and quantum computing. We are driven to discover. We are home to 3,000+ researchers including 5 Nobel Laureates, 9 US National Medals of Technology, 5 US National Medals of Science, 6 Turing Awards and 13 Inductees in the National Inventors Hall of Fame.

PricewaterhouseCoopers

https://www.pwc.com/

PwC Analytics works with clients to identify new opportunities and realize value that arise from highly available data, low-cost computing technology, and advances in information theory. We investigate topics across all aspects of the analytics 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.

Tulane University

http://tulane.edu

Tulane University is a private, nonsectarian research university in New Orleans, Louisiana, United States. Tulane is among the top national universities in the United States, ranked 40th among the best national universities and in the top 25 for its service learning programs. The Office of Academic Affairs and Provost oversees the appointment, advancement, and retention of faculty; the articulation and deployment of all academic and research programs; and, community engagement. In February 2016, the Carol Lavin Bernick Family Foundation, whose generous support of Tulane University built the iconic and award-winning Lavin-Bernick Center for University Life, pledged a generous donation to support Tulane's faculty. The Davis Washington Mitchell Lecture Fund was established by Mrs. Ida Mitchell Looney in memory of her grandfather.

Organizing associations:

Sponsors: