

Preface

This volume records the second conference of the International Association for Safe and Ethical Artificial Intelligence, held at UNESCO House in Paris from 24 to 26 February 2026. It is the first IASEAI conference to produce peer-reviewed proceedings.

IASEAI’26 was built on the conviction that the questions that matter most about contemporary AI do not fit inside any one discipline. Alignment, robustness, security, interpretability, and control are entangled with fairness, accountability, human rights, governance, labor, sustainability, conflict, and democratic deliberation. Progress therefore depends on communities that do not routinely speak to one another developing a shared vocabulary, testing one another’s assumptions, and arriving at a common account of both the problems and the responses available.

That conviction shaped the program. Rather than sorting sessions into the familiar domains —alignment, fairness, policy—we organized them around questions no single field can answer on its own: What does it mean to build a trustworthy AI ecosystem? How should vulnerable and underserved users be protected? What role should AI play in human conflict, and what role in democratic deliberation? Each of the 15 parallel-track sessions drew contributions from academic research, industry practice, public policy, and civil-society work, and each closed with moderated discussion and audience questions, so that disagreement and synthesis belonged to the program rather than to the corridor afterward.

The program comprised 92 research papers, 33 selected talks, five panels, keynote addresses, and a workshop day. The conference received 300 research-paper submissions; all were anonymized and evaluated under double-blind review, and authors received their reviews and could respond before decisions were final. This volume publishes the 57 archival papers in full. Titles and abstracts of the 35 accepted non-archival contributions are available at the stable link <https://iaseai.org/iaseai26>.

In these pages you will find the remarkably diverse range of challenges in AI ethics and safety. We hope the volume serves the community by illuminating a core sample of the vital efforts and approaches being developed to address them.

Mark Nitzberg

IASEAI'26 Conference Chair

Jaime Fernández Fisac and Atoosa Kasirzadeh

IASEAI'26 Program Co-Chairs

Acknowledgements

We thank the authors who entrusted their work to IASEAI’26; the speakers, panelists, moderators, and workshop organizers who shaped the live program; and the participants whose questions and conversations carried that work beyond the stage. We thank the 16 Area Chairs and the more than 200 reviewers who assessed the submissions. Their judgment and domain expertise held a single standard of rigor across very different fields, guarded against conflicts of interest, and set the precedent that future IASEAI proceedings will follow. We thank Stuart Russell, IASEAI President; Loy Sheflott, General Manager; Julia Irwin, Associate Executive Director; and Adam Gleave, Workshop Chair, who designed and ran the workshop day. We thank the IASEAI staff —Markela Zeneli, Livia Morris, and J.P. Gonzales —and the 15 IASEAI volunteers who kept three days of logistics running. We thank Elinor Camlin, Maxi Gering, Louise Lilja, and Sara Molina Rodríguez of Camlin Events, who spent months planning and delivering the conference alongside the Conference Committee and our hosts at UNESCO. And we thank the UNESCO team for opening UNESCO House to us and for the partnership that made an international gathering of this kind possible, in particular Iulia Nechifor, Guilherme Godoi, and Davide Storti. IASEAI’26 was made possible by the generous support of the Future of Life Institute.

About IASEAI

Established in October 2024, IASEAI is a global nonprofit organization with a member-elected president and council dedicated to ensuring that AI systems operate safely and ethically, for the benefit of all humanity. IASEAI connects mission-aligned individuals and groups with expertise in multiple relevant areas: technical AI safety, policy and governance, civil society, philosophical and applied ethics, and more. We aim to foster collaboration, provide public resources and content, and boost the work of more than 160 affiliate organizations and 2,500 members.

Copyright and Publication

Published by The AAAI Press, Washington, DC, USA. Copyright © 2026, Association for the Advancement of Artificial Intelligence. All rights reserved.

IASEAI'26 Organizing Committee

Led by

Interim Executive Director & Conference Chair

Mark Nitzberg

University of California, Berkeley; Co-Founder, Center for Human-Compatible AI

President pro tem

Stuart Russell

University of California, Berkeley

Program Chairs

Atoosa Kasirzadeh

Assistant Professor of Philosophy, Carnegie Mellon University; AI researcher

Jaime Fernandez Fisac

Assistant Professor of Electrical and Computer Engineering, Princeton University

Workshop Chair

Adam Gleave

Founder & CEO, FAR.AI

IASEAI'26 Program Committee —Area Chairs

Alex Glaser

Professor in the School of Public and International Affairs and in the Department of Mechan-

ical and Aerospace Engineering, Princeton University

Andrea Bajcsy

Assistant Professor in the Robotics Institute and School of Computer Science, Carnegie Mellon University

Aylin Caliskan

Associate Professor, University of Washington Paul G. Allen School of Computer Science & Engineering

Bharat Chandar

Postdoctoral Researcher, Stanford Digital Economy Lab and Institute for Human-Centered Artificial Intelligence

Dylan Hadfield-Menell

Associate Professor of EECS, MIT

Fazl Barez

Senior Research Fellow in Technical AI Safety and Governance, University of Oxford

Hadas Kress-Gazit

Geoffrey S.M. Hedrick Sr. Professor at the Sibley School of Mechanical and Aerospace Engineering, Cornell University; Associate Dean for Diversity and Academic Affairs, Cornell's College of Engineering

Jake Metcalf

PhD Senior Researcher, Data & Society

Jonathan Stray

Senior Scientist, UC Berkeley Center for Human-Compatible AI (CHAI)

Joshua Kroll

Assistant Professor of Computer Science, Naval Postgraduate School in Monterey, California

Lewis Hammond

DPhil Candidate in Computer Science, University of Oxford; Co-Director, Cooperative AI Foundation

Luc Rocher

Associate Professor, UKRI Future Leaders Fellow; Senior Research Fellow, Oxford Internet Institute

Noam Kolt

Assistant Professor in the Faculty of Law and in the School of Computer Science and Engineering, Hebrew University

Robert Trager

Director of the Oxford Martin AI Governance Initiative, University of Oxford; Senior Research Fellow, Oxford's Blavatnik School of Government

Roel Dobbe

Assistant Professor in Technology, Policy, and Management, Delft University of Technology

Ting Dong

Associate Professor, Tsinghua University