ALGORITHMIC FAIRNESS ACROSS ALIGNMENT PROCEDURES AND AGENTIC SYSTEMS

1 MOTIVATION AND POSITIONING

AI has transitioned from predictive models to interactive, autonomous agents capable of reasoning, planning, and executing complex goals. As the systems increasingly influence social, economic, and scientific decisions, they determine whose interests are represented and whose opportunities are constrained. Ensuring fairness, therefore, is no longer an ethical preference but a practical imperative.

AI systems are no longer limited to making isolated predictions. They now reason, interact, and generate content across text, images, and other modalities. As a result, the fairness challenges they pose have fundamentally transformed (Liu et al., 2023; Vázquez & Garrido-Merchán, 2024; Lin & Losavio, 2025; Afreen et al., 2025). Traditional fairness frameworks, developed primarily for prediction and/or prediction-based decision-making (e.g., classification, regression, clustering tasks) no longer suffice. The canonical algorithmic fairness focus is increasingly inadequate for addressing the procedural (Tang et al., 2024; Xiao et al., 2025), temporal (Liu et al., 2018; Zhang et al., 2020; Tang et al., 2023a; Gupta et al., 2025), and emergent (Zhao et al., 2018; Parrish et al., 2021; Shavit et al., 2023; Li et al., 2025; Wang et al., 2025) fairness issues that arise with advanced AI systems. In particular, the AI community faces an urgent question:

How do fairness principles/tools evolve when AI systems not only predict, but also adapt and act?

This question can take concrete form in a range of practical contexts, each revealing distinct facets of the fairness challenge, for instance:

Practical Example 1 : During reinforcement learning from human feedback (RLHF), preference datasets may encode cultural or gender biases in "helpful" responses. We need to incorporate fairness at the preference modeling stage, not only in final outputs.

Practical Example 2 : In finance, a multi-agent trading system can be trained for profit maximization. Even if each agent adheres to fairness constraints, their collective dynamics might systematically disadvantage smaller market participants.

Practical Example 3 : In healthcare, debiasing a model to remove correlations between gender and disease risk may inadvertently erase clinically relevant information, reducing diagnostic accuracy for women.

This workshop, **Algorithmic Fairness Across Alignment Procedures and Agentic Systems (AFAA)**, emerges at this pivotal moment as a timely forum for rethinking fairness in AI alignment processes and agentic system development.

This workshop directly aligns with several key focus areas in the ICLR 2026 Call for Workshops, including societal consideration and generative models. By examining fairness across alignment procedures and agentic systems, this workshop creates a crucial forum for bridging the gap between rapid technical advances in model capabilities and the equally important advances needed in frameworks of algorithmic fairness to govern these powerful systems.

1.1 PROBLEM STATEMENT AND SIGNIFICANCE

Our workshop addresses critical gaps at the intersection of **Algorithmic Fairness** (Loftus et al., 2018; Corbett-Davies & Goel, 2018; Mitchell et al., 2018; Narayanan, 2018; Makhlouf et al., 2020; Mehrabi et al., 2021; Zhang & Liu, 2021; Pessach & Shmueli, 2022; Tang et al., 2023b) and the rapidly evolving landscape of AI systems, with specific emphasis on **AI Alignment Procedures** (Leike et al., 2018; Bai et al., 2022; Ouyang et al., 2022; Rafailov et al., 2023; Ji et al., 2023; Poddar et al., 2024; Bhaskar et al., 2025; Carichon et al., 2025) and **Agentic AI Systems** (Park et al.,

2023; Kasirzadeh & Gabriel, 2025; Acharya et al., 2025; Piao et al., 2025; Fang et al., 2025). This workshop aims to bring together researchers to explore how fairness principles must evolve to meet the challenges posed by AI's expanding capabilities and deployment contexts.

Emerging Challenge I: Algorithmic Fairness in AI Alignment Procedures

Example research questions:

- What are the procedural requirements for an AI system to be considered "aligned" with human values on fairness and justice, beyond algorithmic fairness metrics that are based on substantive outcomes?
- What types of stereotypical shortcuts do AI systems learn to exploit in their intermediate steps, which may not be visible in input-output analyses, that result in covert discriminative behaviors deviating from the goal of value alignment?
- How to formulate intermediate but more tractable targets during alignment procedures, such that the goal of achieving fairness and justice can be performed with more explicit procedural supervisions?

Why Emergent Challenge I Matters? *Algorithmic Fairness in AI Alignment Procedures* represents a significant shift in how we conceptualize trustworthy AI development (e.g., addressing Practical Example 1). AI systems increasingly operate in high-stakes domains, for instance, scientific discovery (Jumper et al., 2021; Pyzer-Knapp et al., 2022; Szymanski et al., 2023; Abramson et al., 2024), healthcare (Asan et al., 2020; Bajwa et al., 2021; Alowais et al., 2023; Olawade et al., 2024), legal practice (Shaver, 2023; Zhong, 2023; Lifshitz & Hung, 2024; Merken, 2025), and so on. As systems grow more complex, their decision-making processes become increasingly opaque to human understanding. Ensuring alignment with human values thus requires examining not just outcomes but the legitimacy of the processes behind them (Leike et al., 2018; Bai et al., 2022; Ouyang et al., 2022; Rafailov et al., 2023; Ji et al., 2023; Poddar et al., 2024; Bhaskar et al., 2025; Carichon et al., 2025). Current approaches tend to overlook procedural fairness, allowing systems to appear just while violating fairness principles in disguise (Tang et al., 2024). Developing intermediate, interpretable targets for procedural supervision can make alignment with fairness and justice principles more transparent and controllable. This direction is crucial for building AI systems that are both effective and trustworthy.

Emerging Challenge II: Algorithmic Fairness in Agentic AI Systems

Example research questions:

- What are short-term and long-term fairness implications when AI agents make sequences of interdependent decisions over extended time horizons, where early choices potentially constrain future options for different groups?
- How do we ensure fairness when multiple AI agents interact in competitive and/or collaborative settings, potentially creating emergent discriminatory dynamics that no single agent intended or was able to prevent effectively on its own?
- How do we design procedural fairness constraints that are robust to agents' context-adaptive behaviors while protecting diverse stakeholder interests?

Why Emergent Challenge II Matters? *Algorithmic Fairness in Agentic AI Systems* introduces unprecedented temporal and interactive dimensions to fairness research (e.g., addressing Practical Example 2). As AI systems evolve from single-decision tools to autonomous agents operating over extended periods, they create complex temporal dynamics (OpenAI, 2024; DeepSeek-AI, 2025; Anthropic, 2025; Mistral, 2025). Accordingly, fairness must be evaluated not just at individual decision points, but across entire trajectories of interaction. Persistent AI agents accumulate history and relationships that can create long-term advantages or disadvantages for certain groups. When multiple agents interact in shared environments like markets or social platforms, emergent behaviors can arise that no single agent's fairness constraints can prevent (Park et al., 2023; Kasirzadeh & Gabriel, 2025; Acharya et al., 2025; Piao et al., 2025; Fang et al., 2025). Their adaptive learning processes further complicate fairness, as responsiveness to context may undermine broader consis-

tency. Ensuring fairness thus requires a procedural and structural perspective that accounts for how autonomous systems shape environments and opportunities over time.

Emerging Challenge III: Algorithmic Fairness and Foundation Models

Example research questions:

- What are short-term and long-term fairness implications when AI agents make sequences of interdependent decisions over extended time horizons, where early choices potentially constrain future options for different groups?
- How do we debias foundation models without degrading their general capabilities or destroying useful and unbiased knowledge about demographic differences?
- What are the trade-offs between pre-training debiasing, fine-tuning interventions, and inference-time bias mitigation strategies?

Why Emergent Challenge III Matters? Toward a broader scope, foundation models present a unique paradox for fairness research: their massive scale and broad capabilities make them both the most important targets for debiasing efforts and the most challenging to modify without unintended consequences (e.g., addressing Practical Example 3). The sheer scope of knowledge encoded in these models means that bias mitigation should remove harmful stereotypes while preserving legitimate knowledge about demographic differences that may be crucial for downstream applications like healthcare or social science research (Zhao et al., 2018; Parrish et al., 2021; Tamkin et al., 2023; Li et al., 2025; Wang et al., 2025).

Remark: Emergent Challenges are Deeply Connected These three challenges are deeply interconnected. Progress in fairness-aware alignment procedures (Emergent Challenge I) directly shapes how agentic systems reason and act (Emergent Challenge II), since the values and objectives embedded during alignment determine their in-action behaviors. Conversely, studying fairness in agentic settings (Emergent Challenge II) exposes limitations of existing alignment frameworks (Emergent Challenge I), revealing where algorithmic fairness must account for dynamic, multi-agent interactions. At the same time, both depend critically on foundation models (Emergent Challenge III), whose representations and biases propagate through alignment and agency alike. Therefore, addressing any one of these dimensions in isolation is therefore insufficient. A coherent framework for fairness must consider how foundational architectures, alignment mechanisms, and agentic behaviors co-evolve and mutually benefit one another.

1.2 Novelty and Differentiation

Alignment, in the forms of value alignment (Metz, 2021; Osoba et al., 2020; Umbrello & Van de Poel, 2021), has been a prominent area of research in algorithmic fairness. With the rapid advancement of generative AI models and their expanding capabilities, alignment has not only acquired new meanings but has also begun to exert fresh influence on the literature on algorithmic fairness. This creates an opportunity to revisit existing conversations while using them to inform emerging discussions on AI alignment procedures, reasoning models, agentic systems, and their implications on algorithmic fairness. Our workshop examines these connections between fairness, alignment, and agentic systems, highlighting ongoing progress and identifying emerging challenges.

Points of Difference. Algorithmic fairness, alignment, and agentic systems are three topics that have been explored separately with limited consideration at the intersection. In particular, various related ICLR workshops and other top ML conference workshops have focused on only one of these dimensions at a time. In more specialized venues, algorithmic fairness research has also continued to develop largely in isolation without direct connections to AI alignment procedures or agentic systems. In comparison, our proposed workshop focuses on the critical intersection where algorithmic fairness considerations have implications on, and are reshaped by, advancements across alignment procedures and agentic systems. Attendees will gain a deeper understanding of how algorithmic fairness can inform alignment and agentic system design, how algorithmic fairness needs to adapt to these fundamentally new challenges, access novel frameworks that integrate these domains, and build connections across communities that rarely engage directly.

List of related workshops at ICLR,

- Workshops on Alignment Procedures or Agentic Systems, **but without** a discussion of Algorithmic Fairness:
  - (ICLR 2025) Workshop on Reasoning and Planning for Large Language Models; Workshop on Bidirectional Human-AI Alignment; Second Workshop on Representational Alignment
  - (ICLR 2024) First Workshop on Representational Alignment; Workshop on Large Language Models for Agents
- Workshops on Responsible AI and Foundation Models, **but without** the focus on Alignment Procedures and/or Agentic Systems:
  - (ICLR 2025) Building Trust in LLMs and LLM Applications
  - (ICLR 2024) Workshop on Reliable and Responsible Foundation Models; Secure and Trustworthy Large Language Models
  - (ICLR 2023) Trustworthy and Reliable Large-Scale Machine Learning Models

List of related workshops at other top ML conferences, and related specialized venues,

- (Workshops at other ML conferences) Bridging Language, Agent, and World Models for Reasoning and Planning @ NeurIPS 2025, Foundations of Reasoning in Language Models @ NeurIPS 2025, Reliable and Responsible Foundation Models @ ICML 2025, 2nd Workshop on Models of Human Feedback for AI Alignment @ ICML 2025, The First Workshop on the Application of LLM Explainability to Reasoning and Planning @ COLM 2025, Workshop on AI Agents: Capabilities and Safety @ COLM 2025
- (Algorithmic fairness venues) ACM Conference on Fairness, Accountability, and Transparency (FAccT), AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)

The first five editions of this workshop series at NeurIPS (AFME 2024, AFT 2023, AFCP 2022, AFCR 2021, AFCI 2020) have established it as a venue for crystallizing important problems and directions in algorithmic fairness. Notably, the early editions were among the first to emphasize the intersection of fairness and causality with interpretability, robustness, and privacy. Today, we see a similar inflection point emerging with LLMs, where fairness issues are increasingly intertwined with questions of alignment, reasoning, and generalization. Our workshop aims to actively engage with these intersections, attracting researchers from computer science, philosophy, law, and social sciences due to its theoretical and practical implications.

2 Technical Program

2.1 Workshop Format and Schedule

Building on our experience from earlier versions of the workshop, we will have four invited talks (35 minutes each), five spotlight presentations (5 minutes each), two poster sessions (all posters will be presented in both sessions), three roundtable discussions (in parallel), and one panel discussion.

Two separate poster sessions allow ample time for discussion between participants, and were received positively by authors in the past editions of the workshop. Thus, we plan to maintain this format. Likewise, given the success of roundtable discussions in previous editions, we will continue to include them, focusing on three key roundtable topics mirroring the previously discussed emerging challenges: (a) Algorithmic Fairness in AI Alignment Procedures, (b) Algorithmic Fairness in Agentic AI Systems, and (c) Algorithmic Fairness and Foundation Models. Finally, given the focus of our workshop at the intersection of algorithmic fairness and alignment procedures, we will host a panel discussion featuring experts from both communities to foster cross-disciplinary dialogue.

A tentative schedule of the workshop is shown in Table 1.

2.2 Tentative List of Invited Speakers and Panelists

Table 2 shows the invited speakers, panelists, and roundtable leads. 3 (out of 4) speakers, 2 (out of 5) panelists, and all roundtable leads are confirmed at this point.

Morning	
09:00 - 09:05	Opening Remarks
09:05 - 09:40	Invited Talk
Break	
09:45 - 10:20	Invited Talk
Break	
10:25 - 10:35	Spotlight Presentations
10:35 - 11:20	Poster Session I
Break	
11:25 - 12:00	Invited Talk
Break	

Afternoon	
12:10 - 01:00	Roundtables
Lunch Break	
02:00 - 02:35	Invited Talk
Break	
02:40 - 02:55	Spotlight Presentations
02:55 - 03:55	Poster Session II
Break	
04:15 - 04:55	Panel
Break	
05:00 - 05:10	Closing Remarks

Table 1: Tentative Schedule.

Name	Affiliation	Topic	Status
Invited Speakers			
Mark Riedl	Georgia Institute of Technology	Fairness, Alignment and AI Agents	Confirmed
Stephen Pfohl	Google Research	Fairness, Causality and Foundation Models	Confirmed
Alex Beutel	OpenAI	Fairness and Alignment	Invited
Hima Lakkaraju	Harvard University	Fairness, Interpretability and Foundation Models	Confirmed
Invited Panelists			
Yang Liu	UC Santa Cruz	Algorithmic Fairness	Confirmed
Hanna Wallach	Microsoft Research	Algorithmic Fairness	Invited
Esin Durmus	Anthropic	Alignment Procedures and Agentic Systems	Invited
Atoosa Kasirzadeh	Carnegie Mellon University	Alignment Procedures and Agentic Systems	Invited
Mark Riedl	Georgia Institute of Technology	Alignment Procedures and Agentic Systems	Confirmed
Invited Roundtable Leads			
Francielle Vargas	São Paulo State University	Algorithmic Fairness in AI Alignment Procedures	Confirmed
Yatong Chen	Max Planck Institute	Algorithmic Fairness in Agentic AI Systems	Confirmed
Olawale Salaudeen	Massachusetts Institute of Technology	Algorithmic Fairness and Foundation Models	Confirmed

Table 2: Speakers, panelists, and roundtable leads.

Diversity in speakers, panelists, and roundtable leads. Invited speakers, panelists, and roundtable leads come from different fields (ML, statistics, computational linguistics, ethics, AI safety) and reflect a wide range of perspectives. The group includes researchers from both industry and academia (4 and 8, respectively), and geographically, spans three continents: North America, South America, and Europe. The invitees also reflect varying levels of seniority, with experienced researchers featured in talks and panels, and early-career researchers highlighted as roundtable leads.

2.3 Submission Tracks and Review Process

Main Paper Track To encourage discussion, we will accept submissions between 4 to 9 pages in length of novel work in the area of fairness with a special interest on (but not limited to): normative foundations of alignment procedures; value learning for fairness and alignment; cultural and contextual dimensions of alignment; governance and oversight for agentic systems; ethical challenges and accountability frameworks for agentic systems; long-term societal impacts; defining and measuring fairness for agentic systems; bias mitigation in foundation models; trade-offs between fairness, alignment, and performance in foundation models.

Tiny Paper Track We will also have a 2-page Tiny Paper track to encourage submission of preliminary work, and make the workshop more accessible to potential authors outside the ML conference publication circuit. Accepted submissions from this track will be presented as posters only.

We will use OpenReview for managing submissions, taking advantage of the templates provided by ICLR. We will also use our existing pool of reviewers from previous editions of the workshop. Conflicts of interest will be handled comprehensively, extending beyond institutional overlaps to

include co-authorship, recent collaborations, advisor–advisee relationships, and other professional ties. This ensures a transparent review process aligned with ICLR's standards.

We will be following the policies on Large Language Model Usage at ICLR 2026.

2.4 PROGRAM COMMITTEE

In all previous editions of the workshop, we diligently ensured a minimum of 3 high-quality reviews per submission, with most submissions receiving 4 or more reviews. We were able to accomplish this by (1) utilizing an increasing reviewer pool from previous editions, (2) inviting new reviewers with expertise in the current theme of interest, and (3) sharing an open call for reviewers, providing reviewing experience for new researchers. We carefully matched less experienced reviewers with more experienced researchers in the paper matching step. The program committee in the past edition included 89 reviewers and 9 meta-reviewers. See Table 3 for the 2024 program committee. We plan to utilize a similar process for this year's workshop.

META-REVIEWERS
Ana-Andreea Stoica
Babak Salimi
Christoph Kern
Jessica Schrouff
Kun Zhang
Laurent Charlin
Mattia Cerrato
Stephen R Pfohl
Xueru Zhang

REVIEWERS			
Sri Sri Perangur	Ramya Srinivasan	Mattia Cerrato	Seamus Somerstep
Jan Ramon	Elette Boyle	Aliasghar Khani	Sanne Vrijenhoek
Minyechil Alehegn Tefera	Zeyu Tang	Eike Petersen	Xuchen Li
Elliot Creager	Hadis Anahideh	Isacco Beretta	Mina Arzaghi
Agoritsa Polyzou	Maarten Buyl	Adrián Arnaiz-Rodríguez	Isabela Albuquerque
Megha Srivastava	Julien Ferry	Melissa Hall	Vidhya Kamakshi
Sanghamitra Dutta	Chen Liang	Abdelrahman Zayed	Dimitri Staufer
Aleksander Wieczorek	Samuel Dooley	Stacey Truex	Federico Peiretti
Kate Donahue	Otto Sahlgren	Daniela Cialfi	Esubalew Desta Asmare
Prakhar Ganesh	Marianne Abemgnigni Njifon	Martin Lopatka	Gökhan Özbülek
Vishal Bhalla	Zairah Mustahsan	Matthew Landers	Sukanya Moorthy
Anoush Najarian	Amin Nikanjam	Marta Marchiori Manerba	Tareen Dawood
Jonas Ngnawe	Ranya Aloufi	Ishmeet Kaur	Peeyush Agarwal
Ziqing Yang	Debashis Ghosh	MaryBeth Defrance	Arian Khorasani
Taofeek Abayomi	Canyu Chen	Georgia Baltsou	Prasanjit Dubey
Rakshit Naidu	Yanan Long	Afaf Taik	Jiahao Li
Krystal Maughan	Aparna Balagopalan	Shenao Yan	Rajeev Ranjan Dwivedi
Kamorudeen A Amuda	Tim Räz	Muhammad Mohsin	Kimon Kieslich
Alan Mishler	Saber Malekmohammadi	Sebastian Zezulka	David Hartmann
Robin Burke	Martina Cinquini	Deborah D Kanubala	Shomik Jain
Jan Simson	Samuel R Mayworm	Zhiyu Guo	Sofia Jaime
Andrés Domínguez Hernández	David Kinney	Haolun Wu	Christine Herlihy
			Matteo Fabbri

Table 3: Program committee of our previous workshop edition (AFME 2024).

2.5 ANTICIPATED AUDIENCE SIZE

The previous edition of our workshop attracted 150+ active in-person participants at NeurIPS 2024. We have observed an increasing trend in both workshop attendance and submission throughout the years. Given the high interest in the topics of the workshop and previous years' attendance record, we estimate similar figures this year.

2.6 TIMELINE

Our proposed timeline is as follows:

- Abstract deadline: January 23, 2026 (Anywhere on Earth)
- Submission deadline: January 30, 2026 (Anywhere on Earth)
- Deadline for reviews: February 20, 2026 (Anywhere on Earth)
- Deadline for meta-reviews: February 25, 2026 (Anywhere on Earth)
- Author Notification: March 1, 2026 (Anywhere on Earth)

3 Accessibility and Dissemination

3.1 Virtual Participation and Exceptional Circumstances

We will stream all talks and the panel to the online audience, using the streaming platform provided by ICLR. We will also set up an online meeting to allow virtual presentations or talks only in exceptional circumstances, such as visa issues or other constraints.

3.2 Knowledge Dissemination and Persistence

We maintain a website for all editions of the workshop (https://www.afciworkshop.org), and will use the same website to host accepted papers. Additionally, we have also been publishing selected papers from this workshop series in the PMLR proceedings for the last three years, and plan to continue doing the same for the current version of the workshop.

3.3 Junior Researchers Focus

We provide several opportunities to highlight junior researchers in our workshop. The Tiny Paper track allows junior researchers to get feedback on early-stage work, and the spotlight talks are aimed at further highlighting promising work from the community. Supporting the authors by publishing high-quality work at PMLR was also appreciated in previous editions, and we will continue to do the same for this version of the workshop. Finally, the roundtable discussions in previous editions have provided a platform for both junior and senior researchers to engage in meaningful conversations.

Given the positive reception of our workshop, we also intend to seek external funding to provide financial support to in-person presenters and attendees. Additionally, we will ensure that all attendees are aware of funding opportunities (e.g., through groups such as WiML or Black in AI). In previous years, most complementary tickets for workshop organizers were distributed to authors and awarded to attendees based on financial need. We plan to have similar opportunities this year.

4 Organizing Team and Logistics

All members of the core organizing team have substantial organizational experience. Four out of six organizers have organized previous iterations of the workshop at NeurIPS. Two additional organizers are new to organizing this workshop but have previous experience being in the organizing committee of other conferences (CLeaR 2025) and giving tutorials (FAccT 2025, AAAI 2026).

Our advisory members bring deep expertise in both algorithmic fairness and AI alignment, and will provide strategic guidance on program development to help ensure the workshop maintains rigorous standards while fostering productive dialogue between communities.

4.1 Core Organizing Members

Zeyu Tang, Postdoctoral Scholar at Stanford University (Stanford, United States), <zeyu@cs.stanford.edu>: Zeyu is a postdoctoral scholar in Computer Science at Stanford University, working on trustworthy and responsible AI. His research focuses on causal learning and reasoning to enhance the capabilities of intelligent systems, as well as on algorithmic fairness to model and understand the societal impacts of computational technologies. During his Ph.D. at Carnegie Mellon University, Zeyu was supported by the National Institute of Justice (NIJ) Graduate Research Fellowship and was named the K&L Gates Presidential Fellow in Ethics and Computational Technologies. He serves on the organizing committee of the 4th Conference on Causal Learning and Reasoning (CLeaR 2025), is a program committee member for ICLR, NeurIPS, ICML, FAccT, CLeaR, UAI, AAAI, AISTATS, CVPR, ICCV, and ICDM, and helped organize the *A-General-I Reading Group* and the *Foundation Model Principled Way Reading Group* at CMU.

Prakhar Ganesh, Ph.D. student at McGill University and Mila (Montreal, Canada), <prakhar.ganesh@mila.quebec>: Prakhar is a Ph.D. student at McGill University and Mila. His research spans several aspects of multiplicity with an emphasis on fairness, and

has been recognized with the Best Paper Award at FAccT 2023 and a spotlight presentation at AFME@NeurIPS 2024. He is a recipient of the prestigious FRQNT Doctoral Training Scholarship Award 2024-2028, the McGill Graduate Excellence Award 2024, and the Mila Excellence Scholarship for EDI in research 2024-2027. Prakhar is one of the organizers of the tutorial, "The Many Faces of Multiplicity in ML," presented at ACM FAccT 2025 (and to be also presented at AAAI 2026), a program committee member of ICLR, NeurIPS, ICML, FAccT, ARR, AAAI, AISTATS, WACV, and has been a volunteer for FAccT 2024, ICLR 2025, and FAccT 2025.

Awa Dieng, Ph.D. student at Massachusetts Institute of Technology (Cambridge, USA), <awadieng@mit.edu>: Awa is a Ph.D. student at the Massachussetts Institute of Technology. Her research focuses on building reliable machine learning systems that can be deployed safely, with a particular interest in identifying and understanding sources of bias to inform effective mitigation strategies. Her work is supported by the MIT Presidential Graduate Fellowship. Awa was an organizer of the AFME2024, AFT 2023, AFCP 2022, AFCR 2021, AFCI 2020 NeurIPS workshops, a program chair of the Montreal AI Symposium 2022, a volunteer at ICML 2020 and has served as a reviewer for several ML venues including ICLR, NeurIPS, ICML, AISTATS, and TMLR.

Miriam Rateike, Research Scientist at IBM Research Africa (Nairobi, Kenya), Ph.D. student at Saarland University (Saarbrücken, Germany), <miriam.rateike@ibm.com>: Miriam is a Research Scientist at IBM Research Africa, and a Ph.D. student at Saarland University and an ELLIS student. Her Ph.D. research focuses on algorithmic fairness and feedback loops. She is also enrolled in legal studies and thus particularly interested in the intersection of fairness and law. Miriam received the Google AI Fellowship 2023 in Machine Learning. She is an organizer of the TrustAI 2025 workshop at Deep Learning Indaba, and was an organizer of four NeurIPS workshops AFME 2024, AFT 2023, AFCP 2022, AFCR 2021, as well as the ELLIS workshop on Causethial Machine Learning 2021, and the TReND Python Course 2022, 2021.

Jamelle Watson-Daniels, Research Scientist at Meta FAIR (Atlanta, United States), <watsondaniels@meta.com>: Jamelle is a Research Scientist at Meta FAIR. She completed her Ph.D. in applied math at Harvard. Her interdisciplinary research interests span a few areas: algorithmic fairness, predictive reliability & robustness, ethical & social implications of algorithms. Jamelle has been awarded the Ford Foundation Predoctoral Fellowship and NSF Graduate Research Fellowship. She has experience conducting research at Google and Microsoft. And before pursuing her Ph.D., she received a combined degree in Physics and Africana Studies from Brown University. She was an organizer of a previous NeurIPS workshop AFME 2024 and has served on the programming committee several ML conferences including NeurIPS, ICML, FAccT, AAAI and ICLR.

Golnoosh Farnadi, Assistant Professor at McGill University and Mila (Montreal, Canada), <farnadig@mila.quebec>: Golnoosh is an assistant professor at the school of computer science at McGill University and a core academic member at Mila (Quebec AI Institute). She is also a faculty researcher at Google. In 2021, Golnoosh was appointed a Canada AI CIFAR chair for her work on algorithmic fairness in AI. Golnoosh is a recipient of the 2021 Google Research Scholar Award, the 2021 Facebook Research Award on privacy-preserving technology, and 2023 Google Inclusion award. She was named one of the 2022 Rising Stars, finalist of WAI 2023 responsible AI leader of the year, and one of the 100 Brilliant Women in AI Ethics in 2023. Golnoosh was one of the organizers of the 1st Mila/IVADO summer school on Bias and Discrimination in 2018 in Montreal and has been the scientific director of the online MOOC based on its content. She was an organizer of the AFME 2024, AFT 2023, AFCP 2022, AFCR 2021, AFCI 2020 NeurIPS workshop, the NeurIPS 2022 tutorial on *algorithmic fairness at the intersection* and the senior program chair of the Montreal AI Symposium 2022. She also co-organized the 1st Responsible Generative AI Workshop at CVPR2024, and she was a co-tutorial chair of ACM FAccT 2025.

4.2 Advisory Members

Jessica Schrouff, Director of Responsible AI at GlaxoSmithKline (GSK) (London, United Kingdom), <jvp.schrouff@gmail.com>: Jessica is a Director of Responsible AI at GlaxoSmithKline (GSK) in London, UK. Previously, she was a Senior Research Scientist at Google DeepMind, where she worked at the intersection of machine learning (ML) and healthcare, with

a special interest in responsible ML. Her current research investigates how techniques for explainability, robustness, fairness and causality can lead to more credible machine learning models for healthcare. Jessica received her Ph.D. in Electrical Engineering from the University of Liège, Belgium, in 2013. She was then a post-doctoral researcher at Stanford University and a Marie Curie fellow Research Associate at University College London. She has organized multiple events, including the Pattern Recognition for NeuroImaging (PRNI) workshop at Stanford in 2015 (Steering Committee 2016–2018), multiple PRoNTo educational courses, the AFCP 2022, AFCR 2021, AFCI 2020 NeurIPS workshop, and the Machine Learning for Healthcare (ML4H) symposium 2021.

Sanmi Koyejo Assistant Professor at Stanford University (Stanford, United States), <sanmi@cs.stanford.edu>: Sanmi is an Assistant Professor in the Department of Computer Science at Stanford University and an adjunct Associate Professor at the University of Illinois at Urbana-Champaign. He leads the Stanford Trustworthy AI Research (STAIR) lab, which develops measurement-theoretic foundations for trustworthy AI systems, spanning AI evaluation science, algorithmic accountability, and privacy-preserving machine learning, with applications to healthcare and scientific discovery. His research on AI capabilities evaluation has challenged conventional understanding in the field, including work on measurement frameworks cited in the 2024 Economic Report of the President. Sanmi has received the Presidential Early Career Award for Scientists and Engineers (PECASE), Skip Ellis Early Career Award, Alfred P. Sloan Research Fellowship, NSF CAREER Award, and multiple outstanding paper awards at flagship venues, including NeurIPS and ACL. He has delivered keynote presentations at major conferences, including ECCV and FAccT. He serves in key leadership roles, including Board President of Black in AI, Board of Directors of the Neural Information Processing Systems Foundation, and other leadership positions in professional organizations advancing AI research and broadening participation in the field.

4.3 Diversity in organizers

The core organizing members include a diverse set of researchers, reflecting a range of perspectives and diversity in gender, race, and cultural background. Geographically, the organizers come from institutions on 2 continents and 3 countries: Africa (Kenya) and North America (Canada, USA). Additionally, the team spans a wide range of seniority levels from both academia and industry, including 3 Ph.D. students, a postdoctoral researcher, a research scientist, and an assistant professor.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024.

Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. Agentic AI: Autonomous intelligence for complex goals–a comprehensive survey. *IEEE Access*, 2025.

Juveria Afreen, Mahsa Mohaghegh, and Maryam Doborjeh. Systematic literature review on bias mitigation in generative ai. *AI and Ethics*, pp. 1–53, 2025.

Shuroug A Alowais, Sahar S Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I Alshaya, Sumaya N Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A Badreldin, et al. Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education*, 23(1):689, 2023.

Anthropic. System card: Claude Opus 4 & Claude Sonnet 4. 2025.

Onur Asan, Alparslan Emrah Bayrak, and Avishek Choudhury. Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research*, 22(6):e15154, 2020.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Junaid Bajwa, Usman Munir, Aditya Nori, and Bryan Williams. Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthcare Journal*, 8(2):e188–e194, 2021.

Adithya Bhaskar, Xi Ye, and Danqi Chen. Language models that think, chat better. *arXiv preprint arXiv:2509.20357*, 2025.

Florian Carichon, Aditi Khandelwal, Marylou Fauchard, and Golnoosh Farnadi. The coming crisis of multi-agent misalignment: Ai alignment must be a dynamic and social process. *arXiv preprint arXiv:2506.01080*, 2025.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jinyuan Fang, Yanwen Peng, Xi Zhang, Yingxu Wang, Xinhao Yi, Guibin Zhang, Yi Xu, Bin Wu, Siwei Liu, Zihao Li, et al. A comprehensive survey of self-evolving ai agents: A new paradigm bridging foundation models and lifelong agentic systems. *arXiv preprint arXiv:2508.07407*, 2025.

Kishor Datta Gupta, Mohd Ariful Haque, Hasmot Ali, Marufa Kamal, Syed Bahauddin Alam, and Mohammad Ashiqur Rahman. Continuous monitoring of large-scale generative ai via deterministic knowledge graph structures. *arXiv preprint arXiv:2509.03857*, 2025.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Atoosa Kasirzadeh and Iason Gabriel. Characterizing AI agents for alignment and governance. *arXiv preprint arXiv:2504.21848*, 2025.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. Prompting fairness: Integrating causality to debias large language models. In *Proceedings of the 13th International Conference on Learning Representations*, 2025.

Lisa R Lifshitz and Roland Hung. BC Tribunal confirms companies remain liable for information provided by AI chatbot. https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-february/bc-tribunal-confirms-companies-remain-liable-information-provided-ai-chatbot/, 2024.

Xiaojian Lin and Michael Losavio. A comprehensive survey on bias and fairness in generative ai: Legal, ethical, and technical responses. *Ethical, and Technical Responses (March 04, 2025)*, 2025.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Sara Merken. AI 'hallucinations' in court papers spell trouble for lawyers. https://www.reuters.com/technology/artificial-intelligence/ai-hallucinations-court-papers-spell-trouble-lawyers-2025-02-18/, 2025.

Thaddeus Metz. African reasons why ai should not maximize utility. In *African values, ethics, and technology: Questions, issues, and approaches*, pp. 55–72. Springer, 2021.

Mistral. Magistral. *arXiv preprint arXiv:2506.10910*, 2025.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, volume 1170, pp. 3, 2018.

David B Olawade, Aanuoluwapo C David-Olawade, Ojima Z Wada, Akinsola J Asaolu, Temitope Adereni, and Jonathan Ling. Artificial intelligence in healthcare delivery: Prospects and pitfalls. *Journal of Medicine, Surgery, and Public Health*, 3:100108, 2024.

OpenAI. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Osonde A Osoba, Benjamin Boudreaux, and Douglas Yeung. Steps towards value-aligned systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 332–336, 2020.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*, 2021.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52516–52544, 2024.

Edward O Pyzer-Knapp, Jed W Pitera, Peter WJ Staar, Seiji Takeda, Teodoro Laino, Daniel P Sanders, James Sexton, John R Smith, and Alessandro Curioni. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, 8(1):84, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Nicola Shaver. The use of large language models in LegalTech. https://www.legaltechnologyhub.com/contents/the-use-of-large-language-models-in-legaltech/, 2023.

Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. Practices for governing agentic ai systems. *Research Paper, OpenAI*, 2023.

Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91, 2023.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*, 2023.

Zeyu Tang, Yatong Chen, Yang Liu, and Kun Zhang. Tier Balancing: Towards dynamic fairness over underlying causal factors. In *International Conference on Learning Representations*, 2023a.

Zeyu Tang, Jiji Zhang, and Kun Zhang. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Computing Surveys*, 55(13s):1–37, 2023b. ISSN 0360-0300.

Zeyu Tang, Jialu Wang, Yang Liu, Peter Spirtes, and Kun Zhang. Procedural fairness through decoupling objectionable data generating components. In *International Conference on Learning Representations*, 2024.

Steven Umbrello and Ibo Van de Poel. Mapping value sensitive design onto ai for social good principles. *AI and Ethics*, 1(3):283–296, 2021.

Adriana Fernández de Caleya Vázquez and Eduardo C Garrido-Merchán. A taxonomy of the biases of the images created by generative artificial intelligence. *arXiv preprint arXiv:2407.01556*, 2024.

Angelina Wang, Michelle Phan, Daniel E Ho, and Sanmi Koyejo. Fairness through difference awareness: Measuring desired group discrimination in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6867–6893. Association for Computational Linguistics, 2025.

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *Journal of the American Statistical Association*, pp. 1–21, 2025.

Xueru Zhang and Mingyan Liu. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*, pp. 525–555. Springer, 2021.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? In *Advances in Neural Information Processing Systems*, volume 33, pp. 18457–18469, 2020.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, 2018.

Monica Zhong. The impact of LLMs on the legal industry. https://www.edgeworthecono mics.com/insight-impact-LLMs-legal-industry, 2023.