XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge

Gianmarco Mengaldo, Jiawen Wei, Christopher J. Anders, Emtiyaz Khan, Abeba Birhane, Sara Hooker

1. Workshop Summary

Abstract: Machine learning (ML) models are impressive when they work but they can also show unreliable, untrustworthy, and harmful dangerous behavior. Such behavior is even more common in the era of large models, such as chatGPT, which are quickly being adopted even though we do not understand why they work so well and fail miserably at times. Unfortunately, such rapid dissemination encourages irresponsible use, for example, to spread misinformation or create deep fakes, while hindering the efforts to use them to solve pressing societal problems and advance human knowledge. Ideally, we want models that have a human-like capacity to learn by observing, theorizing, and validating the theories to improve the understanding of the world. At the very least, we want them to aid human knowledge and help us to further enrich it.

Our goal in this workshop is to bring together researchers working on understanding model behavior and show how this key aspect can lead to discovering new human knowledge[1]. The workshop will include theoretical topics on understanding model behavior, namely interpretability and explainability[2,3,4,5] (XAI), but also three distinct scientific application areas: weather and climate, healthcare, and material science (ML4Science)[6]. These three topics are brought together to highlight how seemingly diverse applied scientific fields can leverage XAI for knowledge discovery. The list of topics on theory (Tx) and practice (Px) is given below

- (T1) A-priori (i.e., ante-hoc) interpretability and self-explainable models
- (T2) A-posteriori (i.e., post-hoc) interpretability and attribution methods
- (P1) XAI for knowledge discovery in weather and climate applications
- (P2) XAI for knowledge discovery in the healthcare sector
- (P3) XAI for knowledge discovery in material science

Our organizers, speakers, and panelists have expertise working on such topics and they come from a diverse group in terms of backgrounds, fields, affiliations, gender, race, and seniority. We also have an advisory committee who have been helping us in organization and communication. Some of the organizers have experience in organizing similar events at top ML conferences. There have been a number of workshops on interpretability and explainability in recent years, at NeurIPS, ICML, and ICLR[7,8,9,10]. This workshop aims to go further by not only focusing on the general goal of understanding and shaping model behavior, but also to facilitate its use in practice to advance human knowledge. We believe this is a timely and important topic that will benefit the ICLR community.

2. Invited Speakers and Panelists

The workshop includes 8 invited talks, 2 panel discussions, and 2 poster sessions (1 hour each). **Below, we give a list of speakers and panelists.** The topic which they cover is shown in parenthesis, for example, (T2,P3), and their status as early career researchers (ECRs) are indicated as well.

1. **(Confirmed)** Cynthia Rudin *Speaker and Panelist* (T1, T2, P1).
2. **(Confirmed)** Marinka Zitnik [ECR] *Speaker and Panelist* (T1, T2, P1).
3. **(Confirmed)** Hanna Wallach *Speaker and Panelist* (T1,T2, P1).
4. **(Confirmed)** Erik Cambria *Speaker and Panelist* (T1, T2, P1).
5. **(Confirmed)** Simon See / Jeff Adie *Speaker and Panelist* (T1, P2).
6. **(Confirmed)** Paris Pedrikaris [ECR] *Speaker and Panelist* (T1, T2, P2).
7. **(Confirmed)** Xavier Bresson *Speaker and Panelist* (T1, T2, P2).
8. **(Confirmed)** Qianxiao Li [ECR] *Speaker and Panelist* (T2, P3).

3. Tentative Schedule

Morning Sessions	
8:15 - 8:30	Introduction by organizers
8:30 - 09:00	Invited talk I (Cynthya Rudin, Duke)
9:00 - 09:30	Invited talk II (Marinka Zitnik, Harvard)
09:30 - 10:30	Coffee + Posters
10:30 - 11:00	Invited talk III (Hanna Wallach, Microsoft)
11:00 - 11:30	Invited talk IV (Erik Cambria, NTU)
11:30 - 12:15	Panel discussion
12:15 - 13:30	Lunch break
Afternoon Sessions	
13:30 - 14:00	Invited talk V (Simon See / Jeff Adie, NVIDIA)
14:00 - 14:30	Invited talk VI (Paris Pedrikaris, UPenn / Microsoft)
14:30 - 15:30	Coffee + Posters
15:30 - 16:00	Invited talk VII (Xavier Bresson, NUS)
16:00 - 16:30	Invited talk VIII (Qianxiao Li, NUS)
16:30 - 17:15	Panel discussion
17:15 - 17:30	Closing remarks

4. Diversity Commitment

Our speakers, panelists, and organizers are a diverse group of individuals in terms of backgrounds, fields, affiliations, gender, race and seniority. We have 6 organizers including researchers of different nationalities (Italy, Germany, France/Luxembourg, India, Ethiopia, China, USA) and at different stages of their career in industry and academia. We have 3 female and 3 male organizers. The organizers are affiliated to institutions from different regions (Asia, US, and Europe). Similarly to the organizers, we have made significant effort to choose speakers and panelists with diverse backgrounds and also to cover the list of topics and complement the organizers' expertise, covering all topics from T1-T2 and P1-P3.

The goal of the workshop is also to bring together a diverse set of researchers working in different fields. The workshop will be advertised through mailing lists such as LXAI, MusIML, Black in AI, and Queer in AI to increase the reach of the call for papers. **We will provide financial support to attendees who need them; this will be supported through non-industry funding made available by the organizers.** All invited talks will be recorded, captioned, and made accessible to ICLR attendees to accommodate all time zones.

5. Modality and Accessibility

The workshop will be held in hybrid format so that it can be accessible to people who cannot attend in person. The workshop will be live-streamed, and all talks and panel discussion will be recorded and posted online. Posters will be displayed on the workshop website, with the ability for remote participants to ask questions asynchronously on Rocket.Chat. We also plan to create a Slack channel for the participants and organizers.

6. Anticipated Audience Size

We envision attracting approximately 100 in-person attendees and 40 to 50 submissions. The attendees will be a mix of workshop organizers, committee members, invited speakers, invited researchers with accepted submissions. We openly invite anyone with an interest towards XAI for knowledge discovery and related topics to attend the workshop. There is no maximum number of attendees other than the capacity of the venue.

We welcome **anonymous** submissions of ongoing and unpublished work on any topics related to the workshop, including but not limited to the listed topics (T1, T2) and (P1, P2, P3). Both authors and reviewers will be anonymous throughout the reviewing process. Each paper will receive at least two reviews. For the sake of a smooth reviewing process, all submissions must be made via **OpenReview**. Submissions must be in a **single** PDF file and are required to use the ICLR 2025 LaTeX template. The main paper is limited to **four content pages**, including all figures and tables. Unlimited pages are allowed for references and appendices in the same PDF as the main paper.

7. Previous Related Workshops

This is the first edition of the workshop XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge. Previous workshops on related topics include: Physics for Machine Learning (ICLR 2023), AI for Scientific Discovery: From Theory to

Practice (NeurIPS 2023), XAI in Action: Past, Present, and Future Applications (NeurIPS 2023), eXplainable AI approaches for debugging and diagnosis (NeurIPS 2021). In contrast to previous workshops, this workshop goes further by not only focusing on the general goal of understanding model behavior, but also to facilitate its use in practice to advance human knowledge in the applied sciences. The workshop aims to bring together researchers from various fields to foster interdisciplinary discussions, and enable the flow of knowledge between machine learning and scientific communities.

8. Organizers and Biographies

We have 6 organizers including researchers of different nationalities (Italy, Germany, France/Luxembourg, India, Ethiopia, China, USA) and at different stages of their career in industry and academia. We have 3 female and 3 male organizers. The organizers are affiliated to institutions from different regions (Asia, US, and Europe), and are committed to foster collaboration and networking among researchers across the globe who are working on this topic.

Below are the biographies and details of the organizers. The designated-contact persons are: Gianmarco Mengaldo, Jiawen Wei.

Gianmarco Mengaldo [scholar] [email] is an Assistant Professor in the Department of Mechanical Engineering at National University of Singapore (Singapore), and an Honorary Research Fellow at Imperial College London (United Kingdom). He received his BSc and MSc in Aerospace Engineering from Politecnico di Milano (Italy), and his PhD in Aeronautical Engineering from Imperial College London (United Kingdom). After his PhD he undertook various roles both in industry and academia, including at the European Centre for Medium-Range Weather Forecasts (ECMWF), the California Institute of Technology (Caltech), and Keefe, Bruyette and Woods (KBW). Gianmarco adopts an interdisciplinary approach integrating mathematical and computational engineering to study complex systems that arise in applied science. His current research interests involve (i) explainable AI, both theoretical and applied, (ii) the intersection between AI and domain knowledge, (iii) high-fidelity multi-physics simulation tools, and (iv) data-mining technologies for coherent pattern identification. Dr Mengaldo's main application areas include engineering, geophysics, healthcare, and finance. Dr Mengaldo has extensive experience in organizing workshops in international conferences (he organized 8 workshops), and industry (2 workshops).

Jiawen Wei [scholar][email] is a PhD student in the Department of Mechanical Engineering at National University of Singapore, advised by Dr. Gianmarco Mengaldo. Currently, she is working on trustworthy machine learning for time series, and her research interests include neural network interpretability, interpretability evaluation, explainable artificial intelligence (XAI), time series pattern discovery, and XAI for science. She has past experience in helping with workshop organization at the university (2 workshops).

Christopher J. Anders [scholar] [email] is a research associate at the Berlin Institute for the Foundations of Learning and Data and a post-doctoral researcher in the Machine Learning Group at Technische Universität Berlin. His research encompasses explainable machine learning, specifically the identification and mitigation of model bias and the robustness of feature attribution methods, software for machine learning, and machine learning for physical

sciences (lattice field theory and variational quantum eigensolver). He finished his PhD in machine learning at Technische Universität Berlin in 2024.

Emtiyaz Khan [scholar] [email] is a team leader at the RIKEN Center for Advanced Intelligence Project (AIP) in Tokyo where he leads the Approximate Bayesian Inference Team. The main goal of his research is to understand the principles of learning from data and use them to develop algorithms that can learn like living beings. He has past experience serving as Program chair of ICLR 2024 and Workshop Chair for ICML 2020 and ACML 2021, among numerous other organizing experiences at all top ML conferences. He also organized an ICML 2023 workshop on “Duality Principles for Modern ML”. He serves in the advisory committee for NeurIPS 2022 workshop on "Gaussian Processes, spatiotemporal modeling and Decision making". He has been a reviewer for NeurIPS workshops. He also has experience organizing similar events at many other occasions, for example, Dagstuhl meeting on AI for Social Good in 2019 and 2023, and Fields’ Institute Conference on Data Science.

Abeba Birhane [scholar][email] is currently a Senior Advisor in Trustworthy AI at Mozilla Foundation. She is also an Adjunct Lecturer/Assistant Professor at the School of Computer Science and Statistics at Trinity College Dublin, Ireland. She is a cognitive scientist researching human behavior, social systems, and responsible and ethical Artificial Intelligence (AI). She recently finished her PhD, where she explored the challenges and pitfalls of automating human behavior through critical examination of existing computational models and audits of large scale datasets. Birhane featured on the TIME100 Most Influential People in AI list and serves on the United Nations Secretary-General’s AI Advisory Body.

Sara Hooker [scholar][email] leads Cohere For AI, a non-profit research lab that seeks to solve complex machine learning problems. Cohere For AI supports fundamental research that explores the unknown, and is focused on creating more points of entry into machine learning research. With a long track-record of impactful research at Google Brain, Sara brings a wealth of knowledge from across machine learning. Her work has focused on model efficiency training techniques and optimizing for models that fulfill multiple desired criteria -- interpretable, efficient, fair and robust. Before Cohere For AI, she was the founder of Delta Analytics, a non-profit that brings together researchers, data scientists, and software engineers to volunteer their skills for non-profits around the world.

9. Programme Committee

Our program committee consists of seven researchers who are leading experts on various topics encompassing model understanding and machine learning for the sciences. Below is the list of the committee members **(all confirmed)**.

1. **(Confirmed)** Wojciech Samek
2. **(Confirmed)** Klaus-Robert Müller
3. **(Confirmed)** Sebastian Lapuschkin
4. **(Confirmed)** Prasanna Balaprakash
5. **(Confirmed)** Hugues Turbé

List of References

[1] G. Mengaldo. **Explain the Black Box for the Sake of Science: the Scientific Method in the Era of Generative Artificial Intelligence**, ArXiv https://arxiv.org/abs/2406.10557 (2024).

[2] C. Rudin. **Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead**, Nature Machine Intelligence (2019).

[3] D. Alvarez Melis, T. Jaakkola. **Towards robust interpretability with self-explaining neural networks**, *Advances in neural information processing systems* 31 (2018).

[4] https://www.nytimes.com/2018/12/26/science/chess-artificial-intelligence.html

[5] M. Raissi, P. Perdikaris, G.E. Karniadakis. **Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations**, Journal of Computational Physics, 378, pages 686-307, 2019.

[6] H. Wang et al. **Scientific discovery in the age of artificial intelligence**, Nature 620, pages 47–60, 2023.

[7] http://interpretable-ml.org/icml2020workshop/

[8] https://xai-in-action.github.io/

[9] https://xai4debugging.github.io/

[10] https://sites.google.com/view/trustml-unlimited/home