# ICML-21 Workshop on
# Theoretic Foundation, Criticism, and Application Trend of Explainable AI

Ying Nian Wu[1], Quanshi Zhang[2], Tian Han[3], Zhanxing Zhu[4], Lixin Fan[5], and Hang Su[6]
[1] University of California, Los Angeles [2] Shanghai Jiao Tong University
[3] Stevens Institute of Technology [4] Huawei Technologies [5] WeBank AI [6] Tsinghua University

**1. Workshop summary:** Deep neural networks (DNNs) have undoubtedly brought great success to a wide range of applications in computer vision, computational linguistics, and AI. However, foundational principles underlying the DNNs' success and their resilience to adversarial attacks are still largely missing. Interpreting and theorizing the internal mechanisms of DNNs becomes a compelling yet controversial topic.

Unlike previous workshops or tutorials on explainable AI (XAI), the proposed workshop pays a special interests in **theoretic foundations**, **limitations**, and **new application trends** in the scope of XAI. These issues reflect new bottlenecks in the future development of XAI, for example: (1) no theoretic definition of XAI and no solid and widely-used formulation for even a specific explanation task. (2) No sophisticated formulation of the essence of "semantics" encoded in a DNN. (3) How to bridge the gap between connectionism and symbolism in AI research has not been sophisticatedly explored. (4) How to evaluate the correctness and trustworthiness of an explanation result is still an open problem. (5) How to bridge the intuitive explanation (*e.g.,* the attribution/importance-based explanation) and a DNN's representation capacity (*e.g.,* the generalization power) is still a significant challenge. (6) Using the explanation to guide the architecture design or substantially boost the performance of a DNN is a bottleneck.

Therefore, this workshop aims to bring together researchers, engineers as well as industrial practitioners, who concern about the interpretability, safety, and reliability of artificial intelligence. In this workshop, we hope to use a broad discussion on the above bottleneck issues to explore new critical and constructive views of the future development of XAI. Research outcomes are also expected to profoundly influences critical industrial applications such as medical diagnosis, finance, and autonomous driving.

**2. Topics:** Topics of interests include, but are not limited to, the following fields:
- XAI theories
- Critical and constructive commentary on XAI, *e.g.,* limitations of the current XAI techniques.
- Qualitative and quantitative diagnosis and analysis of the XAI systems.
- Deep coupling of neural networks and grammars or graphical models
- Probabilistic logic programming, causality reasoning and learning.
- Safety, adversarial attacking and defense of DNNs.
- Industrial applications of trustworthy AI, *e.g.* in medical diagnosis, autonomous driving, and finance.

All above topics are core issues in the development of explainable AI and have received increasing attention in recent years. We believe the workshop would be of broad interest to the ICML community.

**3. Tentative program schedule:** The workshop will include eight invited talks (25 minutes + 5 minutes QA), five contributed talks (13 minutes + 2 minutes QA), "poster" sessions (5 mins videos) and a panel discussion.

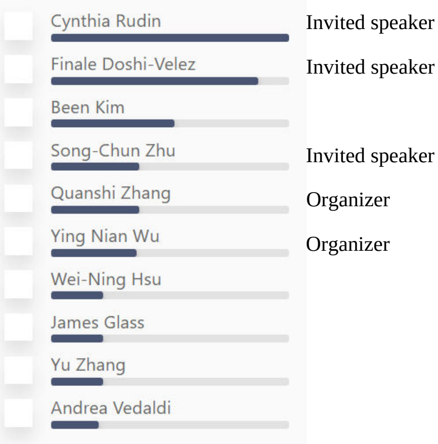
The tentative schedule is listed below (in EDT, IT: invited talk; CT: contributed talk; PS: poster session; ):

| 02:50 | 03:00 | 04:00 | 04:30 | 05:30 | 06:30 | 07:30 | 08:00 | 09:00 | 09:30 | 10:00 | 11:00 | 12:00 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Welcome | IT 1/2 | CT 1/2 | PS1 | IT 3/4 | Break | CT 3/4 | IT 5/6 | CT 5 | Break | IT 7/8 | PS2 | **Panel** |

**4. Related workshops and tutorials in history:** We list four related workshops and tutorials as follows.
- ICML workshop on XXAI: Extending Explainable AI Beyond Deep Models and Classifiers, 2020
- CVPR Tutorial on Interpretable Machine Learning for Computer Vision, 2020
- AAAI workshop on Network Interpretability for Deep Learning, 2019
- CVPR workshop on Explainable AI, 2019

Compared to the above relevant workshops on XAI, the proposed workshop aims to covers more diverse topics, becaues this workshop pays special interests **theoretic foundations**, **limitations**, **new application trends**, and **new explanation direction** of XAI. The workshop welcomes scientists, engineers, practitioners in both academic and industrial communities who are interested in the interpretability, safety, and reliance of artificial intelligence.

**5. Speakers:**



| | |
|---|---|
| Cynthia Rudin | Invited speaker |
| Finale Doshi-Velez | Invited speaker |
| Been Kim | |
| Song-Chun Zhu | Invited speaker |
| Quanshi Zhang | Organizer |
| Ying Nian Wu | Organizer |
| Wei-Ning Hsu | |
| James Glass | |
| Yu Zhang | |
| Andrea Vedaldi | |

- **Dr. Song-Chun Zhu** (confirmed), Professor, UCLA, **IEEE Fellow**, Dean of Institute of AI, Peking University, China
- **Dr. Finale Doshi-Velez** (confirmed), Associate Professor, Harvard University, USA
- **Dr. Yan Liu** (confirmed), Associate Professor, University of Southern California, USA
- **Dr. Cynthia Rudin** (confirmed), ASA Fellow, Professor, Duke University, USA
- **Dr. Mukund Sundararajan** (confirmed), Principal research scientist/director at Google, USA
- **Dr. Klaus-robert Muller**, Professor, TU-berlin, Germany
- **Dr. Su-In Lee**, Associate Professor, University of Washington, USA
- **Dr. Jun Zhu**, Professor, Tsinghua University, China

The above figure shows top-10 authors ranked by *Microsoft Academic Search* (https://academic.microsoft.com/home) with the keyword *interpretable* in the time range between 2017 and 2020. **Among the top-6 scholars, two scholars are organizers of this workshop, and three scholars are the invited speakers.**

**6. Make the workshop interactive:** We aim to hold a one-day workshop remotely via Zoom. The Zoom session will be live-streamed and recorded for audiences in difference time zones.

**7. Organizers:**
- **Ying Nian Wu**, Professor, University of California, Los Angeles
- **Quanshi Zhang**, Associate Professor, Shanghai Jiao Tong University
- **Tian Han**, Assistant Professor, Stevens Institute of Technology
- **Zhanxing Zhu**, Senior Scientist, Huawei Technologies
- **Lixin Fan**, Principal Scientist, Department of AI at WeBank, China
- **Hang Su**, Associate Professor, Tsinghua University, China

*Diversity commitment:* Most organizers and all speakers are top-6 scholars, who are ranked by the Microsoft Academic Search with the keyword "interpretable" in the time range from 2017 to 2020. Furthermore, all organizers and speakers mainly focus on different issues within the scope of interpretability, including attribution/importance-based explanation, the explanation of information processing logic in DNNs, the explanation of the representation power of DNNs, etc. Speakers achieve gender parity and are also diverse with respect to affiliations, nationalities, and scientific background.

## Appendix: organizer bibliographic references and qualification

- **Ying Nian Wu**, Professor, University of California, Los Angeles, ywu@stat.ucla.edu
  Web:http://www.stat.ucla.edu/~ywu/ [Google Scholar]
- **Quanshi Zhang**, Associate Professor, Shanghai Jiao Tong University, zqs1022@sjtu.edu.cn
  Web: http://qszhang.com [Google Scholar]
- **Tian Han**, Assistant Professor, Stevens Institute of Technology, than6@stevens.edu
  Web: https://hthth0801.github.io [Google Scholar]
- **Zhanxing Zhu**, Senior Scientist, Huawei Technologies zhanxing.zhu@pku.edu.cn
  Web: https://sites.google.com/view/zhanxingzhu/home [Google Scholar]
- **Lixin Fan**, Principal Scientist, Department of AI at WeBank, China, lixinfan@webank.com
  [Google Scholar]
- **Hang Su**, Associate Professor Tsinghua University, China, suhangss@mail.tsinghua.edu.cn
  [Google Scholar]

The organizers have rich experience of organizing the following workshops in top-tier conferences.

- CVPR workshop on Neural Architecture Search and Beyond for Representation Learning, 2020
  https://sites.google.com/view/cvpr20-nas/
- CVPR workshop on Explainable AI, 2019
  https://explainai.net/
- AAAI workshop on Network Interpretability for Deep Learning, 2019
  http://networkinterpretability.org/
- CVPR workshop on Language and Vision, 2017
  http://languageandvision.com/2017.html
- **Dr. Quanshi Zhang** is an associate professor at Shanghai Jiao Tong University. He is ranked as the **Fifth Most Influential Scholar in XAI** by *Microsoft Academic Search* using the keyword *interpretable* within the duration 2017–2020. He has received the **ACM China Rising Star Award 2021** for his contributions to XAI. He is founded by the "Thousand Youth Talents Plan." He gave a panel discussion on XAI in ICML 2020, a tutorial on *trustworthiness of interpretable machine learning* in IJCAI 2020, and gave an invited talk on XAI in AAAI 2019. In the scope of explainable AI, he has published 23 papers in top-tier conferences (including IEEE T-PAMI, ICML, ICLR, CVPR, ICCV, ECCV, AAAI) from 2017 to 2020. He was the chair of the CVPR Workshop on Explainable AI 2019 and the chair of the AAAI Workshop on Network Interpretability for Deep Learning 2018. When he was a researcher at the University of California, Los Angeles, before 2018, he was supported by the grant of the DARPA XAI project (total 5,999,988 USD). Four of his studies in explainable AI were reported by synced review (https://syncedreview.com/). Please see the website qszhang.com for details.
  – The Fifth Most Influential Scholar in XAI by the Microsoft Academic Search using the keyword *interpretable* within the duration 2017–2020 – ACM China Rising Star Award 2021 for his contributions to XAI
  – IJCAI 2020 Tutorial on trustworthiness of interpretable machine learning
  – ICML 2020 Panel Discussion on Baidu AutoDL: Automated and Interpretable Deep Learning
  – AAAI 2019 Workshop on Network Interpretability, an invited talk

– CVPR 2019 Workshop on Explainable AI, Co-chair
– AAAI 2018 Workshop on Network Interpretability for Deep Learning, Co-Chair

- **Dr. Ying Nian Wu** is a Professor at the University of California, Los Angeles. His research interests include statistics, machine learning, and computer vision. He received a Ph.D. degree from the Harvard University in 1996. He was an Assistant Professor at the University of Michigan between 1997 and 1999 and an Assistant Professor at the University of California, Los Angeles between 1999 and 2001. He became an Associate Professor and a Full Professor at the University of California, Los Angeles in 2001 and 2006, respectively. He participates in the DARPA project on Learning and Communicating Explainable Representations for Analytics and Autonomy.

- **Dr. Tian Han** is an assistant professor in the Department of Computer Science in the Schaefer School of Engineering and Science at Sevens Institute of Technology. Prior to joining the Stevens faculty in 2019, he obtained his Ph.D. from the Statistics Department at the University of California, Los Angeles (UCLA), and his M.Phil. from the Computer Science and Engineering Department at the Hong Kong University of Science and Technology (HKUST). He has served as PC members and reviewers in the dominant international conferences, including AAAI, NeurIPS, CVPR, ECCV, ICCV. He is in the area of artificial intelligence (AI) and machine learning, focusing on developing statistical learning methods for probabilistic models, and on building explainable AI system for various applications. He is currently working on developing disentangled and hierarchical generative models towards an explainable AI system.

- **Dr. Zhanxing Zhu** is a Senior Scientist at Huawei Technologies. Before that, he was an assistant professor at School of Mathematical Sciences, Peking University, also affiliated with Center for Data Science, Peking University. He obtained a Ph.D. degree in machine learning from University of Edinburgh in 2016. His research interests cover machine learning and its applications in various domains. Currently, he mainly focuses on deep learning theory, robustness, and interpretability. Particularly, his research on the interplay between adversarial robustness, interpretability and safety of deep learning has attracted much attention in the machine learning community. He was awarded 2019 Alibaba Damo Young Fellow and obtained Best Paper Finalist from the top computer security conference CCS 2018.

- **Dr. Lixin Fan** is the principal scientist affiliated with WeBank AI Department, Shenzhen, China. His research areas of interest include Machine learning and deep learning, Computer vision and pattern recognition, Image and video processing, 3D big data pro-cessing, data visualization and rendering, Augmented and virtual reality, Mobile ubiquitous and pervasive computing and Intelligent human-computer interface. Dr Fan is the (co-)author of more than 60 international journal and conference publications, with more than 6000 citations. His research work included the well recognized Bag of Keypoints method for image categorization. He is the (co-)inventor of dozens of granted and pending patents filed in US, Europe and China. Dr Fan also co-organized workshops on a variety of topics including federated learning, held jointly with conferences such as NeurIPS, CVPR, ICCV, IJCAR, ICPR, ICME and ISMAR.

- **Dr. Hang Su** is an associate professor in the Department of Computer Science and Technology at Tsinghua University. Before joining Tsinghua, he received his Ph. D. degree from Shanghai Jiaotong University in 2014 and worked as a visiting scholar at Carnegie Mellon University from 2011 to 2013. His research interests lie in the development of computer vision and machine learning algorithms with a recent focus on robust and interpretable machine learning algorithms. He has published around 50 papers including CVPR, ECCV, TMI, and served as senior PC or PC members in the dominant international conferences including IJCAI, AAAI, CVPR. He received Young Investigator Award from MICCAI2012, the Best Paper Award in AVSS2012, and Platinum Best Paper Award in ICME 2018.