

2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML 2023)

**Raleigh, North Carolina, USA
8 – 10 February 2023**



**IEEE Catalog Number: CFP23BT6-POD
ISBN: 978-1-6654-6300-3**

**Copyright © 2023 by the Institute of Electrical and Electronics Engineers, Inc.
All Rights Reserved**

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP23BT6-POD
ISBN (Print-On-Demand):	978-1-6654-6300-3
ISBN (Online):	978-1-6654-6299-0

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

CURRAN ASSOCIATES INC.
proceedings
.com

2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) **SaTML 2023**

Table of Contents

Message from the Program Committee Co-Chairs	xi
Organizing Committee	xii
Program Committee	xiii
Steering Committee	xv
Keynote Addresses	xvi
Sponsors	xvii

Fairness

Explainable Global Fairness Verification of Tree-Based Classifiers	1
<i>Stefano Calzavara (Università Ca' Foscari Venezia, Italy), Lorenzo Cazzaro (Università Ca' Foscari Venezia, Italy), Claudio Lucchese (Università Ca' Foscari Venezia, Italy), and Federico Marcuzzi (Università Ca' Foscari Venezia, Italy)</i>	
Exploiting Fairness to Enhance Sensitive Attributes Reconstruction	18
<i>Julien Ferry (LAAS-CNRS, Université de Toulouse, CNRS, France), Ulrich Aïvodji (Ecole de Technologie Supérieure, Canada), Sébastien Gambs (Université du Québec à Montréal, Canada), Marie-José Huguet (LAAS-CNRS, Université de Toulouse, CNRS, INSA, France), and Mohamed Siala (LAAS-CNRS, Université de Toulouse, CNRS, INSA, France)</i>	
Wealth Dynamics Over Generations: Analysis and Interventions	42
<i>Krishna Acharya (Georgia Institute of Technology, USA), Eshwar Ram Arunachaleswaran (University of Pennsylvania, USA), Sampath Kannan (University of Pennsylvania, USA), Aaron Roth (University of Pennsylvania, USA), and Juba Ziani (Georgia Institute of Technology, USA)</i>	
Learning Fair Representations through Uniformly Distributed Sensitive Attributes	58
<i>Patrik Joslin Kenfack (Innopolis University, Russia), Adín Ramírez Rivera (University of Oslo, Norway), Adil Mehmood Khan (Innopolis University, Russia; University of Hull, UK), and Manuel Mazzara (Innopolis University, Russia)</i>	

Privacy

Can Stochastic Gradient Langevin Dynamics Provide Differential Privacy for Deep Learning?	68
<i>Guy Heller (University of Bar-Ilan, Ramat Gan, Israel) and Ethan Fetaya (University of Bar-Ilan, Ramat Gan, Israel)</i>	
Kernel Normalized Convolutional Networks for Privacy-Preserving Machine Learning	107
<i>Reza Nasirigerdeh (Technical University of Munich, Germany), Javad Torkzadehmahani (Azad University of Kerman, Iran), Daniel Rueckert (Technical University of Munich, Germany; Imperial College London, United Kingdom), and Georgios Kaissis (Technical University of Munich, Germany; Helmholtz Zentrum Munich, Germany; Imperial College London, United Kingdom)</i>	
Model Inversion Attack with Least Information and an In-Depth Analysis of its Disparate Vulnerability	119
<i>Sayanton V. Dibbo (Dartmouth College), Dae Lim Chung (Dartmouth College), and Shagufra Mehmaz (The Pennsylvania State University)</i>	
Distribution Inference Risks: Identifying and Mitigating Sources of Leakage	136
<i>Valentin Hartmann (EPFL), Léo Meynent (EPFL), Maxime Peyrard (EPFL), Dimitrios Dimitriadis (Amazon), Shruti Tople (Microsoft Research), and Robert West (EPFL)</i>	
Dissecting Distribution Inference	150
<i>Anshuman Suri (University of Virginia), Yifu Lu (University of Michigan), Yanjin Chen (University of Virginia), and David Evans (University of Virginia)</i>	

Distributed and Collaborative Learning

ExPloit: Extracting Private Labels in Split Learning	165
<i>Sanjay Kariyappa (Georgia Institute of Technology) and Moinuddin K Qureshi (Georgia Institute of Technology)</i>	
SafeNet: The Unreasonable Effectiveness of Ensembles in Private Collaborative Learning	176
<i>Harsh Chaudhari (Northeastern University), Matthew Jagielski (Google Research), and Alina Oprea (Northeastern University)</i>	
Reprogrammable-FL: Improving Utility-Privacy Tradeoff in Federated Learning via Model Reprogramming	197
<i>Huzaifa Arif (Rensselaer Polytechnic Institute, USA), Alex Gittens (Rensselaer Polytechnic Institute, USA), and Pin-Yu Chen (IBM Research, USA)</i>	
Optimal Data Acquisition with Privacy-Aware Agents	210
<i>Rachel Cummings (Columbia University), Hadi Elzayn (Stanford University), Emmanouil Pountourakis (Drexel University), Vasilis Gkatzelis (Drexel University), and Juba Ziani (Georgia Institute of Technology)</i>	

Integrity at Inference

A Light Recipe to Train Robust Vision Transformers	225
<i>Edoardo Debenedetti (ETH Zurich, Switzerland), Vikash Sehwal (Princeton University, USA), and Prateek Mittal (Princeton University, USA)</i>	
Less is More: Dimension Reduction Finds On-Manifold Adversarial Examples in Hard-Label Attacks	254
<i>Washington Garcia (University of Florida), Pin-Yu Chen (IBM Research), Hamilton Clouse (Air Force Research Laboratory), Somesh Jha (University of Wisconsin), and Kevin Butler (University of Florida)</i>	
Publishing Efficient On-Device Models Increases Adversarial Vulnerability	271
<i>Sanghyun Hong (Oregon State University), Nicholas Carlini (Google Brain), and Alexey Kurakin (Google Brain)</i>	
EDoG: Adversarial Edge Detection For Graph Neural Networks	291
<i>Xiaojun Xu (University of Illinois at Urbana-Champaign), Hanzhang Wang (eBay), Alok Lal (eBay), Carl Gunter (University of Illinois at Urbana-Champaign), and Bo Li (University of Illinois at Urbana-Champaign)</i>	
Counterfactual Sentence Generation with Plug-and-Play Perturbation	306
<i>Nishtha Madaan (IBM Research India; Indian Institute of Technology), Diptikalyan Saha (IBM Research India), and Srikanta Bedathur (Indian Institute of Technology)</i>	
Rethinking the Entropy of Instance in Adversarial Training	316
<i>Minseon Kim (KAIST, South Korea), Jihoon Tack (KAIST, South Korea), Jinwoo Shin (KAIST, South Korea), and Sung Ju Hwang (KAIST, South Korea; AITRICS, South Korea)</i>	
Towards Transferable Unrestricted Adversarial Examples with Minimum Changes	327
<i>Fangcheng Liu (Peking University), Chao Zhang (Peking University), and Hongyang Zhang (University of Waterloo)</i>	
"Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice	339
<i>Giovanni Apruzzese (University of Liechtenstein), Hyrum S. Anderson (Robust Intelligence), Savino Dambra (Norton Research Group), David Freeman (Meta), Fabio Pierazzi (King's College London), and Kevin A. Roundy (Norton Research Group)</i>	
What are Effective Labels for Augmented Data? Improving Calibration and Robustness with AutoLabel	365
<i>Yao Qin (Google Research, USA), Xuezhi Wang (Google Research, USA), Balaji Lakshminarayanan (Google Research, USA), Ed H. Chi (Google Research, USA), and Alex Beutel (Google Research, USA)</i>	

Integrity at Training Time

Sniper Backdoor: Single Client Targeted Backdoor Attack in Federated Learning	377
<i>Gorka Abad (Radboud University, The Netherlands; Ikerlan research centre, Spain), Servio Paguada (Radboud University, The Netherlands; Ikerlan research centre, Spain), Oguzhan Ersoy (Radboud University, The Netherlands), Stjepan Picek (Radboud University, The Netherlands), Víctor Julio Ramírez-Durán (Ikerlan research centre, Spain), and Aitor Urbietá (Ikerlan research centre, Spain)</i>	
Backdoor Attacks on Time Series: A Generative Approach	392
<i>Yujing Jiang (University of Melbourne), Xingjun Ma (Fudan University), Sarah Monazam Erfani (University of Melbourne), and James Bailey (University of Melbourne)</i>	
VENOMAVE: Targeted Poisoning Against Speech Recognition	404
<i>Hojjat Aghakhani (University of California, Santa Barbara), Lea Schönherr (CISPA Helmholtz Center for Information Security), Thorsten Eisenhofer (Ruhr University Bochum), Dorothea Kolossa (Technische Universität Berlin), Thorsten Holz (CISPA Helmholtz Center for Information Security), Christopher Kruegel (University of California, Santa Barbara), and Giovanni Vigna (University of California, Santa Barbara)</i>	

Interpretability and Explainability

Endogenous Macrodynamics in Algorithmic Recourse	418
<i>Patrick Altmeyer (Delft University of Technology, The Netherlands), Giovan Angela (Delft University of Technology, The Netherlands), Aleksander Buszydlík (Delft University of Technology, The Netherlands), Karol Dobiczek (Delft University of Technology, The Netherlands), Arie van Deursen (Delft University of Technology, The Netherlands), and Cynthia C. S. Liem (Delft University of Technology, The Netherlands)</i>	
ModelPred: A Framework for Predicting Trained Model from Training Data	432
<i>Yingyan Zeng (Virginia Tech, USA), Jiachen T. Wang (Princeton University, USA), Si Chen (Virginia Tech, USA), Hoang Anh Just (Virginia Tech, USA), Ran Jin (Virginia Tech, USA), and Ruoxi Jia (Virginia Tech, USA)</i>	
Harnessing Prior Knowledge for Explainable Machine Learning: An Overview	450
<i>Katharina Beckh (Fraunhofer IAIS, Germany), Sebastian Müller (University of Bonn, Germany), Matthias Jakobs (TU Dortmund University, Germany), Vanessa Toborek (University of Bonn, Germany), Hanxiao Tan (TU Dortmund University, Germany), Raphael Fischer (TU Dortmund University, Germany), Pascal Welke (University of Bonn, Germany), Sebastian Houben (Hochschule Bonn-Rhein-Sieg, Germany), and Laura von Rueden (Fraunhofer IAIS, Germany)</i>	
Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks	464
<i>Tilman Rauker (n/a), Anson Ho (Epoch), Stephen Casper (MIT CSAIL), and Dylan Hadfield-Menell (MIT CSAIL)</i>	

Verification in Machine Learning

Reducing Certified Regression to Certified Classification for General Poisoning Attacks	484
<i>Zayd Hammoudeh (University of Oregon, USA) and Daniel Lowd (University of Oregon, USA)</i>	
Neural Lower Bounds for Verification	524
<i>Florian Jaeckle (University of Oxford, UK) and M. Pawan Kumar (University of Oxford, UK)</i>	
Toward Certified Robustness Against Real-World Distribution Shifts	537
<i>Haoze Wu (Stanford University, USA), Teruhiro Tagomori (Stanford University, USA; NRI Secure, Japan), Alexander Robey (University of Pennsylvania, USA), Fengjun Yang (University of Pennsylvania, USA), Nikolai Matni (University of Pennsylvania, USA), George Pappas (University of Pennsylvania, USA), Hamed Hassani (University of Pennsylvania, USA), Corina Pasareanu (Carnegie Mellon University, USA), and Clark Barrett (Stanford University, USA)</i>	
CARE: Certifiably Robust Learning with Reasoning via Variational Inference	554
<i>Jiawei Zhang (University of Illinois Urbana-Champaign, USA), Linyi Li (University of Illinois Urbana-Champaign, USA), Ce Zhang (ETH Zürich, Switzerland), and Bo Li (University of Illinois Urbana-Champaign, USA)</i>	
FaShapley: Fast and Approximated Shapley Based Model Pruning Towards Certifiably Robust DNNs	575
<i>Mintong Kang (University of Illinois at Urbana-Champaign), Linyi Li (University of Illinois at Urbana-Champaign), and Bo Li (University of Illinois at Urbana-Champaign)</i>	

Model Governance

PolyKervNets: Activation-Free Neural Networks For Efficient Private Inference	593
<i>Toluwani Aremu (Mohamed Bin Zayed Institute of Artificial Intelligence, UAE) and Karthik Nandakumar (Mohamed Bin Zayed Institute of Artificial Intelligence, UAE)</i>	
Theoretical Limits of Provable Security Against Model Extraction by Efficient Observational Defenses	605
<i>Ari Karchmer (Boston University, USA)</i>	
No Matter How You Slice It: Machine Unlearning with SISA Comes at the Expense of Minority Classes	622
<i>Korbinian Koch (Universität Hamburg, Germany) and Marcus Soll (NORDAKADEMIE gAG Hochschule der Wirtschaft, Germany)</i>	
Data Redaction from Pre-Trained GANs	638
<i>Zhifeng Kong (University of California San Diego, USA) and Kamalika Chaudhuri (University of California San Diego, USA)</i>	

Responsible AI

Tensions Between the Proxies of Human Values in AI	678
<i>Teresa Datta (Arthur), Daniel Nissani (Arthur), Max Cembalest (Arthur), Akash Khanna (Arthur), Haley Massa (Arthur), and John Dickerson (Arthur)</i>	
A Validity Perspective on Evaluating the Justified Use of Data-Driven Decision-Making Algorithms	690
<i>Amanda Coston (Carnegie Mellon University, USA), Anna Kawakami (Carnegie Mellon University, USA), Haiyi Zhu (Carnegie Mellon University, USA), Ken Holstein (Carnegie Mellon University, USA), and Hoda Heidari (Carnegie Mellon University, USA)</i>	
Author Index	705