PROCEEDINGS OF SPIE

Assurance and Security for AI-enabled Systems

Joshua D. Harguess
Nathaniel D. Bastian
Teresa L. Pace
Editors

22–23 April 2024
National Harbor, Maryland, United States

Sponsored and Published by
SPIE

Volume 13054

Proceedings of SPIE 0277-786X, V. 13054

SPIE is an international society advancing an interdisciplinary approach to the science and application of light.

The papers in this volume were part of the technical conference cited on the cover and title page. Papers were selected and subject to review by the editors and conference program committee. Some conference presentations may not be available for publication. Additional papers and presentation recordings may be available online in the SPIE Digital Library at SPIEDigitalLibrary.org.

The papers reflect the work and thoughts of the authors and are published herein as submitted. The publisher is not responsible for the validity of the information or for any outcomes resulting from reliance thereon.

Please use the following format to cite material from these proceedings:
Author(s), "Title of Paper," in *Assurance and Security for AI-enabled Systems*, edited by Joshua D. Harguess, Nathaniel D. Bastian, Teresa L. Pace, Proc. of SPIE 13054, Seven-digit Article CID Number (DD/MM/YYYY); (DOI URL).

ISSN: 0277-786X
ISSN: 1996-756X (electronic)

ISBN: 9781510674264
ISBN: 9781510674271 (electronic)

Published by
SPIE
P.O. Box 10, Bellingham, Washington 98227-0010 USA
Telephone +1 360 676 3290 (Pacific Time)
SPIE.org
Copyright © 2024 Society of Photo-Optical Instrumentation Engineers (SPIE).

Copying of material in this book for internal or personal use, or for the internal or personal use of specific clients, beyond the fair use provisions granted by the U.S. Copyright Law is authorized by SPIE subject to payment of fees. To obtain permission to use and share articles in this volume, visit Copyright Clearance Center at copyright.com. Other copying for republication, resale, advertising or promotion, or any form of systematic or multiple reproduction of any material in this book is prohibited except with permission in writing from the publisher.

Printed in the United States of America by Curran Associates, Inc., under license from SPIE.

Publication of record for individual papers is online in the SPIE Digital Library.

SPIE. DIGITAL LIBRARY
SPIEDigitalLibrary.org

Paper Numbering: A unique citation identifier (CID) number is assigned to each article in the Proceedings of SPIE at the time of publication. Utilization of CIDs allows articles to be fully citable as soon as they are published online, and connects the same identifier to all online and print versions of the publication. SPIE uses a seven-digit CID article numbering system structured as follows:

- The first five digits correspond to the SPIE volume number.
- The last two digits indicate publication order within the volume using a Base 36 numbering system employing both numerals and letters. These two-number sets start with 00, 01, 02, 03, 04, 05, 06, 07, 08, 09, 0A, 0B … 0Z, followed by 10-1Z, 20-2Z, etc. The CID Number appears on each page of the manuscript.

Contents

v *Conference Committee*

ETHICS, FAIRNESS, RISK

13054 03 **On the role of loss-driven systems analysis in assessing AI ethics** [13054-1]

EXPLAINABILITY

13054 05 **Extracting explanations, justification, and uncertainty from black-box deep neural networks** [13054-8]

13054 06 **Bias, explainability, transparency, and trust for AI-enabled military systems** [13054-9]

13054 07 **Towards neuro-symbolic reinforcement learning for trustworthy human-autonomy teaming** [13054-11]

ATTACKS AND DEFENSES

13054 08 **The AI security pyramid of pain** [13054-126]

13054 09 **Bayesian graph representation learning for adversarial patch detection** [13054-4]

13054 0B **An AI red team playbook** [13054-7]

EVALUATION AND METRICS

13054 0C **NRTK: an open source natural robustness toolkit for the evaluation of computer vision models** [13054-12]

13054 0D **Practical guidance for human-machine teaming assurance of AI-enabled systems** [13054-13]

13054 0F **A security analysis of compressed communication in distributed neural networks** [13054-124]

13054 0G **Exploiting explainability for reinforcement learning model assurance** [13054-10]

APPLICATIONS TO RF, LLMS

13054 0J **Demonstrating confidence in RF machine learning systems** [13054-18]

13054 0K **Differential privacy to mathematically secure fine-tuned large language models for linguistic steganography** [13054-19]

CYBER

13054 0M **The relevance, effectiveness, and future prospects of cyber deception implementation within organizations** [13054-20]

13054 0N **Comparison of cognitively-inspired salience and feature importance techniques in intrusion detection datasets** [13054-21]

13054 0O **Security games with malicious adversaries in the clouds: status update** [13054-23]

AI ASSURANCE FOR MDO APPLICATIONS: JOINT SESSION WITH CONFERENCES 13051 AND 13054

13054 0P **Optimized and secured AI for the tactical edge** [13054-25]

13054 0Q **Assured AI reference architecture** [13054-27]

POSTER SESSION

13054 0R **An AI blue team playbook** [13054-26]

DIGITAL POSTER SESSION

13054 0T **Byzantine-robust federated learning performance evaluation via distance-statistical aggregations** [13054-16]