

PROCEEDINGS OF SPIE

Assurance and Security for AI-enabled Systems 2025

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

**14–16 April 2025
Orlando, Florida, United States**

Sponsored and Published by
SPIE

Volume 13476

Proceedings of SPIE 0277-786X, V. 13476

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 SPIDigitalLibrary.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 2025*, edited by Joshua D. Harguess, Nathaniel D. Bastian, Teresa L. Pace, Proc. of SPIE 13476, Seven-digit Article CID Number (DD/MM/YYYY); (DOI URL).

ISSN: 0277-786X

ISSN: 1996-756X (electronic)

ISBN: 9781510687417

ISBN: 9781510687424 (electronic)

Published by

SPIE

P.O. Box 10, Bellingham, Washington 98227-0010 USA

Telephone +1 360 676 3290 (Pacific Time)

SPIE.org

Copyright © 2025 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**

SPIDigitalLibrary.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*

AI ASSURANCE

- 13476 03 **Explaining model robustness: combining saliency maps and natural robustness testing using XAITK and NRTK** [13476-3]
- 13476 04 **Patterns for combining large language models with knowledge bases to improve assurance, performance, and reliability of AI solutions** [13476-25]

CYBERSECURITY

- 13476 05 **Authentic key agreement scheme for blockchain-based smart grid applications** [13476-5]
- 13476 06 **Unified multimodel fusion for precision defense against evasive denial-of-service attacks** [13476-6]
- 13476 07 **Security design for NLIP: a universal protocol for AI-enabled systems** [13476-24]

AI RED TEAMING

- 13476 08 **How private are your chat adapters? Evaluating the privacy of LoRA fine-tuned large language models with membership inference attacks** [13476-8]
- 13476 09 **Prompt engineering for detecting phishing** [13476-10]
- 13476 0A **Adversarial threat vectors and risk mitigation for retrieval-augmented generation systems** [13476-11]
- 13476 0B **Offensive security for AI systems: concepts, practices, and applications** [13476-12]

AI GOVERNANCE I

- 13476 0C **A framework for the assurance of AI-enabled systems** [13476-14]
- 13476 0D **Securing the future of AI: a holistic approach to trust and robustness** [13476-15]

AI AND DATA ASSURANCE

- 13476 OE **Initial measurement of data quality** [13476-19]
- 13476 OF **DataEval: enhancing end-to-end AI development through comprehensive data analysis**
[13476-20]

AI GOVERNANCE II

- 13476 OH **Secure and decentralized digital twin optimization: a blockchain-enabled federated learning approach for IIoT** [13476-21]
- 13476 OI **Guardrails for safe implementations of AI-based services** [13476-23]
- 13476 OJ **Quantifying adversarial risk of multimodal foundation models for military applications**
[13476-18]