*Security and Safety*, Vol. 3, E2024021 (2024)
https://doi.org/10.1051/sands/2024021
sands.edpsciences.org

Security and Safety in Artificial Intelligence

Editorial

OPEN ACCESS

# Preface: Security and safety in artificial intelligence

Hao Zhang[1,2*], Yu-Gang Jiang[3], Claudio Melchiorri[4] and Gerhard Rigoll[5]

[1] *Department of Control Science and Engineering, Tongji University, Shanghai 200092, China*
[2] *Shanghai Institute of Intelligent Science and Technology, Shanghai 201210, China*
[3] *School of Computer Science, Fudan University, Shanghai 200433, China*
[4] *Department of Electrical, Electronic and Information Engineering, University of Bologna, Bologna, 40136, Italy*
[5] *Institute for Human-Machine Communication, Technical University of Munich, Munich, 80539, Germany*

**Citation** Zhang H, Jiang YG, Melchiorri C and Rigoll G. Preface: Security and safety in artificial intelligence. Security and Safety 2024; **3**: E2024021. https://doi.org/10.1051/sands/2024021

This special topic centers on cutting-edge advancements in the security and safety of artificial intelligence (AI), with a focus on critical applications across domains such as autonomous systems, federated learning, and network security. The rapid evolution of AI algorithms, enabled by advances in hardware and software, has led to transformative applications but also revealed significant vulnerabilities and security risks. This topic aims to address these challenges by showcasing research dedicated to enhancing the resilience, reliability, and privacy of AI systems. Highlighting developments in AI security measures, this topic offers researchers and practitioners a platform to share pioneering theoretical and technical insights and propose innovative approaches in AI security. Topics include adaptive defense mechanisms, privacy-preserving techniques, intrusion detection, game-theoretic models, and data-driven solutions, contributing to a more secure foundation for robust AI applications across a range of critical sectors.

This special topic includes 5 papers covering advancements in the safety and security of artificial intelligence, addressing essential challenges in domains such as autonomous systems, federated learning, and network security. These papers contribute significant insights into AI robustness, adaptive defense mechanisms, and the integration of privacy-preserving techniques. The contributions are as follows:

In the survey "Harnessing dynamic heterogeneous redundancy to empower deep learning safety and security [1]," the authors analyze the persistent challenges in deep learning (DL) due to adversarial and backdoor attacks. The survey identifies the root causes of these issues as inherent DL model limitations, termed as Endogenous Safety and Security (ESS) problems, characterized by non-interpretability, non-recognizability, and non-identifiability. To address ESS, the study proposes a Dynamic Heterogeneous Redundant (DHR) architecture to enhance resilience through diversity, validated across diverse applications. The results confirm DHR's superior performance in addressing ESS over conventional DL defense strategies.

In the paper "A requirements model for AI algorithms in functional safety-critical systems with an explainable self-enforcing network [2]," the authors present a self-enforcing network model that enables software developers to evaluate AI systems' compliance with functional safety requirements. By integrating an explainable feature, the model assesses safety integrity levels, offering a developer-focused approach to ensure regulatory compliance for AI in critical systems.

The paper "Safety-critical nonlinear optimal predictive control with adaptive error elimination algorithm for robotic systems [3]" explores a robust control framework for robotic systems under adverse conditions. This study introduces an adaptive error elimination controller combined with a nonlinear optimal predictive control (NOPC) scheme, enhancing robots' secure operation and stability. The

* Corresponding author (email: **zhang_hao@tongji.edu.cn**)

approach also includes a sliding mode controller for handling environmental uncertainties, achieving superior tracking accuracy for complex, safety-critical trajectories in robotic applications.

The paper "VAEFL: Integrating variational autoencoders for privacy preservation and performance retention in federated learning [4]" introduces VAEFL, a federated learning (FL) framework that employs Variational Autoencoders (VAEs) to protect data privacy without sacrificing model performance. VAEFL strategically separates the model into a private encoder and public decoder, preventing data leakage while maintaining high prediction accuracy. The approach excels in privacy-sensitive sectors like finance, balancing data privacy and utility effectively.

In the paper "Robust object detection for autonomous driving based on semi-supervised learning [5]," the authors propose a semi-supervised learning framework to enhance object detection robustness in autonomous vehicles by leveraging unlabeled data. The study builds a baseline using transfer learning and introduces a co-training method and bounding box augmentation to improve resilience against adversarial attacks, demonstrating state-of-the-art robustness in diverse, challenging environments.

This special topic serves as a platform to exchange advancements in AI security and safety, offering theoretical insights and practical solutions. We extend our appreciation to all authors, reviewers, and editors for their valuable contributions to this special topic.

## References

[1] Zhang F, Chen X and Huang W et al. Harnessing dynamic heterogeneous redundancy to empower deep learning safety and security. Security and Safety 2024; **3**: 2024011. https://doi.org/10.1051/sands/2024011

[2] Kluver C, Greisbach A and Kindermann M et al. A requirements model for ai algorithms in functional safety-critical systems with an explainable self-enforcing network. 2024; **3**: 2024020. https://doi.org/10.1051/sands/2024020

[3] Wang H. Safety-critical nonlinear optimal predictive control with adaptive error elimination algorithm for robotic systems. Security and Safety 2024; **3**: 2024016. https://doi.org/10.1051/sands/2024016

[4] Li Z, Liu Y and Li J et al. VAEFL: Integrating variational autoencoders for privacy preservation and performance retention in federated learning. Security and Safety 2024; **3**: 2024005. https://doi.org/10.1051/sands/2024005

[5] Chen WW, Yan J and Huang W et al. Robust object detection for autonomous driving based on semi-supervised learning. Security and Safety 2024; **3**: 2024002. https://doi.org/10.1051/sands/2024002