EDITORIAL

Towards a Secure Future: Security Challenges for Deep Learning, AI, and Foundation Models in the Next Decade

Donghua Jiang[1] **and Jawad Ahmad**[2,*]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China
[2] Cyber Security Center, Prince Mohammad Bin Fahd University, Al-Khobar 31952, Saudi Arabia

The security landscape of modern computing is being reshaped by deep learning, artificial intelligence (AI), and particularly large-scale foundation models. These systems now sit at the core of decision-making pipelines in healthcare, finance, transportation, public services and national infrastructure. As AI transitions from a "model-in-the-lab" to a "system-in-the-world," the threat surface expands from classical cyber risk into a broader, socio-technical arena: malicious data, malicious prompts, malicious users, compromised supply chains and even compromised incentives. In this setting, security is no longer a peripheral add-on to performance. It is a primary design objective.

Despite rapid progress in model capability, the community increasingly recognizes a recurring gap: many evaluation practices emphasize accuracy, efficiency or benchmark-based robustness, yet remain insufficient to characterize real security. Strong performance metrics do not automatically imply resistance to adversarial manipulation, privacy leakage or system-level exploitation. In high-stakes deployments, the critical question is not whether a model performs well under nominal conditions, but whether it remains reliable under targeted, adaptive and resourceful attacks.

This editorial advocates for a more rigorous and security-centric research agenda. One that treats AI security as an end-to-end discipline spanning theory, algorithms, systems and governance. It also calls for a dedicated scholarly focus on large-model security, where new capabilities create new vulnerabilities, and where adversaries can weaponize model behavior on a scale.

Why AI Security Requires a Dedicated, Modern Lens?

AI security differs from traditional security in two important ways. First, the "attack surface" is often statistical rather than purely logical. Subtle perturbations, distribution shifts and data poisoning can produce disproportionate impact without triggering conventional alarms. Second, the defender's artifacts are probabilistic systems trained on internet-scale data, often opaque and frequently updated. Threat models must account for adaptive attackers who can probe models via APIs, exploit long-context behaviors, and extract information through repeated interaction.

Foundation models intensify these challenges. Their broad generality makes them highly valuable and

Submitted: 13 December 2025
Accepted: 21 December 2025
Published: 02 March 2026

Vol. 2, **No.** 2, 2026.
10.62762/TISC.2025.709595

***Corresponding author:**
✉ Jawad Ahmad
jahmad@pmu.edu.sa

Citation
Jiang, D., & Ahmad, J. (2026). Towards a Secure Future: Security Challenges for Deep Learning, AI, and Foundation Models in the Next Decade. *ICCK Transactions on Information Security and Cryptography*, 2(2), 70–72.

highly targetable. A single model may serve millions of users and multiple downstream applications, meaning a successful exploit can propagate widely. Moreover, the model lifecycle (pretraining, alignment, fine-tuning, retrieval augmentation, tool deployment) introduces distinct security assumptions at each stage. Security claims that ignoring this lifecycle risks becoming incomplete or misleading.

Core Research Directions for Deep Learning and Large-Model Security

A credible security agenda for deep learning and foundation models should incorporate at least the following pillars:

- *Adversarial robustness beyond benchmarks*
  Progress requires moving from narrow perturbation models toward adaptive, goal-driven attackers that exploit modality-specific vulnerabilities (vision, speech, text, multimodal). Robustness should be evaluated under realistic constraints (query limits, transfer attacks, physical-world conditions) and against evolving attack strategies.
- *Data integrity and training-time security*
  Data poisoning, backdoors and supply-chain compromise threaten training pipelines and pretrained checkpoints. Research should prioritize detection, provenance, robust training objectives, watermarking of model weights and verifiable dataset curation. For foundation models, even small amounts of malicious influence can have persistent effects.
- *Privacy, leakage and inference-time attacks*
  Membership inference, property inference, inversion and prompt-based extraction are now practical concerns. For large models, privacy is not only about training data; it includes tool outputs, retrieval co-rpora and conversational context. Defenses should be measured against both classical and learning-based inference attacks.
- *Prompt injection, tool misuse and agent security*
  As models gain tool access (browsers, code execution, databases), security becomes a question of capability control: how to prevent untrusted inputs from hijacking system instructions, exfiltrating secrets or causing unsafe actions. This space demands principled policies, sandboxes, least-privilege tool design and robust auditing.
- *Model stealing, extraction and intellectual property protection*
  APIs and open deployment settings enable model extraction and imitation. Research should explore practical watermarking, fingerprinting, query monitoring and legal/operational strategies that complement technical measures.
- *Alignment under adversarial pressure*
  Safety alignment is not static: attackers actively search for jailbreaks, coercion tactics and multi-step manipulation. The community needs reproducible stress tests, red-teaming protocols and evaluation suites that model adversarial intent, not just benign user error.
- *System-level assurance, verification, and governance*
  AI security must include secure deployment practices, incident response, transparency mechanisms and post-deployment monitoring. "Secure-by-design" engineering should be elevated alongside algorithmic innovation.

What Rigor Should Look Like

To advance the field, security research must be held to standards that match the sophistication of modern attackers: (1) Explicit threat models: Clearly state attacker goals, capabilities, and constraints. (2) Adaptive evaluation: Demonstrate resilience against attackers that adapt to defenses, not only fixed bas-elines. (3) Reproducibility and ablation: Provide evidence that security gains are not artifacts of evaluati-on choices. (4) End-to-end realism: Consider the complete pipeline, data, model, retrieval, tools, UI and operations. (5) Negative results and limitations: Security claims should be calibrated; partial defenses should be described as such.

A Call to the Community

The AI security community is uniquely interdisciplinary: it spans cryptography, machine learning, systems engineering, human factors and policy. This diversity is a strength, but it requires a shared commitment to clarity, rigor and honest threat modeling. We encourage research that not only proposes defenses but also delivers convincing evidence that those defenses matter in realistic environments.

We also emphasize the importance of constructive engagement between academia and industry. Many of the most pressing vulnerabilities arise in deployed systems, where constraints, incentives and user

behaviors shape security outcomes. Collaboration grounded in responsible disclosure and ethical experimentation will be essential.

Closing

Deep learning and foundation models will continue to transform society. Whether that transformation strengthens or undermines trust depends on our ability to build systems that are not merely capable, but secure, private and resilient under pressure. We invite the community to treat AI security as a first-class scientific discipline that demands the same seriousness as cryptography and systems security, and that recognizes security as a measurable, testable property rather than an assumption.

We look forward to work that sets new standards for the security of deep learning, AI and large models that advance not only state-of-the-art performance, but also state-of-the-practice trustworthiness.

Data Availability Statement

Not applicable.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

Donghua Jiang is recognized among the world's top 2% scientists according to the Stanford University ranking. His research interests include deep learning-based privacy-preserving technologies, compressive sensing theory and its applications, and chaotic secure communication. Until now, he has published published more than ten academic articles in internationally prestigious journals, including but not limited to IEEE IOTJ, IEEE TCE, CSF, INS, ESWA, SP and NODY. Additionally, social duties he serves include Associate Editor of International Journal of Informa-tion Security and Privacy (IGI Global, Engineering Index, Emerging Science Citation Index) and ICCK Transactions on Information Security and Cryptography (ICCK, Emerging Journal), Guest Editor of IEEE Transactions on Consumer Electronics (IEEE, Science Citation Index) and Entropy (MDPI, Science Citation Index), as well as Reviewer of IEEE TCSVT, IEEE TITS, IEEE TCE, IEEE ESL, ACM TOMM, ACM MM, CSF, SP, etc. (Email: jiangdh8@mail2.sysu.edu.cn)

Jawad Ahmad is a highly experienced teacher with more than 13 years of teaching and research experience in prestigious institutes. He has conducted teaching and research at renowned institutions such as Prince Mohammad Bin Fahd University (KSA), Edinburgh Napier University (UK), Glasgow Caledonian University (UK), and Hongik University (South Korea), among others. He has also served as a supervisor for several PhD, MSc, and undergraduate students, providing guidance and support for their dissertations. He has published in renowned journals including IEEE Transactions, ACM Transactions, Elsevier, and Springer with over 200 research papers and 7000 citations (H-Index 50). For several consecutive years, his name has been included in the world's top 2% scientists in Computer Science. In 2020, he received the endorsement of UK exceptional talent candidate ("Emerging Leader") for pioneering work in the field of Cybersecurity and AI. To date, he has secured research and funding grants of more than £250K in the UK and Norway, etc., as a Principal Investigator (PI) and a Co-Investigator (Co-I). In terms of academic achievements, he has earned a Gold medal for his outstanding performance in MS and a Bronze medal for his achievements in BS. (Email: jahmad@pmu.edu.sa)