亥姆霍茲資訊安全中心 Dr. Mario Fritz 演講
Trustworthy AI and A Cybersecurity Perspective on Large Language Models

文／秦紫頤 研究助理

Mario Fritz 教授是 CISPA Helmholtz 資訊安全中心的成員，同時也是 Saarland University 的名譽教授。在此之前，他曾領導 Max Planck Institute 資訊科學研究所的研究小組。他擁有廣泛的學術背景，包括作為 TPAMI 的副編輯、Trustworthy Federated Data Analytics 計畫的統籌人，以及 100 多篇科學文章的作者，他在 trustworthy AI 的研究方面做出了重要的貢獻。

在演講中，Fritz 博士強調了網路安全和機器學習相互交織的本質，指出它們已經無法分割。這種交織增加了網路安全的複雜性，使我們在使用服務時難以建立信任。解決這些網路安全和可信度的問題變得至關重要，特別是在潛在的崩潰或災難性事件發生之前。他特別提到人工智慧的迅速演進，尤其是 Generative AI 的出現，以及它對隱私、安全和信任的影響。Fritz 博士呼籲在 Generative AI 開發過程中應提供使用者保證，確保新技術符合我們的期望。

演講的前半部分聚焦於網路隱私的問題，特別是機密性的議題。Fritz 博士首先介紹了 membership inference attack，該攻擊試圖從給定的數據中判斷它是否屬於模型的訓練集。攻擊的成功可能導致攻擊者還原模型的訓練數據，對於本來就訓練在含有隱私資訊的數據上的模型造成嚴重安全風險。Fritz 博士接著提到了 differential privacy 的概念，它一種對抗 membership inference attack 的機制，可以通過在模型算法中引入亂數擾動，使攻擊者難以推斷出個體資訊，進而避免數據泄露的風險。此外，他還討論了 evasion attack，這是一種最常見且容易應用的攻擊方式，攻擊者希望通過微調輸入數據來誤導模型的預測。最後，Fritz 博士探討了提升模型強韌性的一些策略，包

括 certification、Lipschitz bounds 和 randomized smoothing，以確保模型有足夠的信心能夠抵禦對抗性的攻擊。

演講的後半部分轉向人工智慧對社會的更廣泛影響，探討了深度偽造檢測和打擊虛假信息等挑戰。在這一部分，演講者指出隨著人工智慧擔任解釋者的角色，信任和誤解的問題變得更加突出，這與前半部分探討的網路安全問題形成呼應。Fritz 博士強調了事實檢查方法的重要性，但也承認評估 Generative AI 生成內容的困難性，尤其是考慮到人為操縱的可能性。此外，他探討了大型語言模型（LLM）對通用和可轉移的對抗攻擊的問題，引發對這些人工智慧代理進行系統的 prompt discussion，用來發現和檢測問題，這是 trustworthy AI 研究重要的議題之一。

演講的最後，Fritz 博士總結時強調了 LLM 應用安全性研究的重要性。他提到各種攻擊形式的潛在威脅，強調在不斷演變的 LLM 生態系統中理解網路安全威脅是必要的。在追求負責任的人工智慧發展過程中，Fritz 博士的見解為研究人員、實踐者和政策制定者提供了重要的指南。

非常感謝 Mario Fritz 博士能夠來到交大演講，同時我們實驗室也很榮幸能在演講前有機會與 Fritz 博士進行經驗交流。在這次交流中，他不僅給予了我們對於已完成和進行中研究提供了許多有益的反饋，這些意見對於我們未來的研究方向提供了清晰的指引，更是激發了我們對於研究的熱忱。

Speech by Dr. Mario Fritz:
Trustworthy AI and A Cybersecurity Perspective on Large Language Models

Professor Mario Fritz is a faculty at the CISPA Helmholtz Center for Information Security and an honorary professor at Saarland University. He previously led a research group at the Max Planck Institute for Informatics. With a diverse academic background, encompassing roles such as the associate editor of the journal "IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)," the coordinator of the Helmholtz project "Trustworthy Federated Data Analytics," and the author of over 100 scientific articles, Dr. Fritz has made significant contributions to research in trustworthy AI.

During the speech, Dr. Fritz underscored the intertwined nature of cybersecurity and machine learning, emphasizing their inseparability. This interconnection increases the complexity of cybersecurity, posing challenges in establishing trust when utilizing services. Resolving these cybersecurity and trust-related issues becomes paramount, particularly before potential breakdowns or catastrophic events occur. Dr. Fritz highlighted the swift evolution of artificial intelligence, notably the emergence of Generative AI, and its impact on privacy, security, and trust. He advocated for implementing user assurances throughout the development of Generative AI to guarantee that emerging technologies align with our expectations.

The initial portion of the speech delves into the realm of network privacy, explicitly addressing the aspect of confidentiality. Dr. Fritz initiated the discussion by presenting the notion of a membership inference attack aimed at discerning whether given data is part of the model's training set. The success of such an attack poses a significant security threat, potentially enabling the attacker to reconstruct the model's training data, which is particularly risky for models trained on data containing private information. Dr. Fritz subsequently discussed differential privacy, a mechanism to thwart membership inference attacks. This is achieved by introducing random perturbations into the model algorithm, complicating the inference of individual information by attackers and mitigating the risk of data leakage. Additionally, he explored evasion attacks, the most prevalent and easily applicable method wherein attackers seek to manipulate the model's predictions by adjusting input data. Lastly, Dr. Fritz delved into various strategies to bolster model robustness, including certification, Lipschitz bounds, and randomized smoothing to ensure the model can withstand adversarial attacks.

The latter portion of the speech shifted toward the broader impact of artificial intelligence on society, addressing challenges such as identifying deepfakes and combating misinformation. During this segment, the speaker highlighted that as artificial intelligence assumes the role of an interpreter, concerns related to trust and misunderstanding become more pronounced, aligning with the earlier examination of cybersecurity issues. Dr. Fritz emphasized the importance of fact-checking methods while acknowledging the complexities in assessing content generated by Generative AI, particularly in light of potential manipulation. Furthermore, he explored the challenges of large language models (LLMs) dealing with general and transferable adversarial attacks, initiating discussions on system prompts for these AI agents to uncover and address issues—an essential aspect of trustworthy AI research.

In his conclusion, Dr. Fritz emphasized the significance of conducting security research on LLM applications. He discussed the possible threats of different attacks and stressed the need to understand network security threats within the ever-evolving LLM ecosystem. Dr. Fritz's insights provide essential guidance for researchers, practitioners, and policymakers involved in the responsible development of artificial intelligence.

We are deeply grateful for Dr. Mario Fritz's speech at National Yang Ming Chiao Tung University, and our laboratory members feel privileged to have had the opportunity to exchange ideas and experiences with Dr. Fritz before the speech. During this interaction, he furnished us with valuable feedback on both completed and ongoing research projects and imparted clear guidance for our future research endeavors. His suggestions not only made a valuable contribution to our academic pursuits but also reignited our enthusiasm for research.