OUKA
The University of Osaka
Institutional Knowledge Archive

Title	A Study on Security and Privacy for AI Supplied by Malicious Model Trainers
Author(s)	岩花, 一輝
Citation	大阪大学, 2026, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/105504
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

様式3

論 文 内 容 の 要 旨

氏 名 （ 岩花 一輝 ）	
論文題名	A Study on Security and Privacy for AI Supplied by Malicious Model Trainers (悪意あるモデル学習者から提供されるAIのセキュリティとプライバシに関する研究)

論文内容の要旨

Recent advances in artificial intelligence (AI) have accelerated its deployment across society, with AI now widely used in domains such as medicine, transportation, and finance. The development and deployment of AI models form an AI supply chain involving multiple entities, including data providers, model trainers, operators, and model users. Within this supply chain, security and privacy risks have become increasingly evident, such as unintended dangerous model behaviors and unintended data leakage to third parties, posing fundamental challenges to the reliability and trustworthiness of AI systems.
This thesis aims to clarify security and privacy threats that arise from malicious model trainers in the AI supply chain, and to contribute to the establishment of a secure and privacy-preserving AI supply chain systems through the analysis of both attack techniques and corresponding defense mechanisms.
First, in Chapter 2, we propose Secure and Private Gradient Computation (SPGC), a defense mechanism against threats to model security and data privacy posed by malicious local model trainers (data providers) in distributed learning settings such as collaborative or federated learning. SPGC is a secure and privacy-preserving framework that integrates Secure Multiparty Computation (MPC) with Differential Privacy (DP). By concealing gradient information via MPC, SPGC reduces the amount of DP noise required, thereby preventing attacks while maintaining high model accuracy.
On the other hand, it remains unclear to what extent a malicious model trainer in a centralized learning setting can mount attacks that induce dangerous model behavior or compromise data privacy.
We therefore investigate the feasibility of both types of attacks when the model trainer behaves maliciously and discuss possible defense strategies.
In Chapter 3, we propose TALPA, a backdoor attack in which a malicious model trainer induces dangerous model behavior while increasing resistance to detection in both the input space and the latent space. Our results demonstrate that, once the model trainer becomes an adversary, it becomes possible to realize backdoor attacks that are significantly more difficult to defend against than conventional approaches. We further suggest that detecting abnormality across classes in the internal representations of the model may provide a promising direction for defenses against such attacks.
In Chapter 4, we propose a data reconstruction attack in which a malicious model trainer trains the model so that its outputs encode information sufficient to reconstruct user inputs. This attack enables the reconstruction of confidential inputs from model outputs even in confidential computing environments that rely on Trusted Execution Environments or Secure Multiparty Computation. We thereby demonstrate that, once the model trainer becomes an adversary, user privacy can be compromised even when input data are kept secret. As a potential mitigation, we show that transforming inputs to increase their diversity can reduce the effectiveness of this attack.
In summary, by analyzing attack and defense mechanisms under malicious model trainers across different learning settings, this thesis establishes a principled foundation for constructing secure and privacy-preserving AI supply chain systems. We believe that the findings of this thesis will support the advancement of trustworthy AI supply chain systems and enhance the understanding of the complex interplay among security, privacy, and the structure of the AI supply chain.

様式 7

論文審査の結果の要旨及び担当者

氏　名　（岩花一輝）	
論文審査担当者	（職）　氏名 主査　教授　猪俣　敦夫 副査　教授　山口　弘純 副査　教授　下西　英之 副査　教授　猿渡　俊介 副査　教授　小泉　祐揮

論文審査の結果の要旨

本博士論文は、サプライチェーンにおける「悪意あるモデル学習者（model trainer）」が引き起こすセキュリティおよびプライバシーの脅威を体系的に明らかにし、それに対抗するための防御手法の確立を目指したものである。昨今、急速に利用が進化を遂げる AI をベースとしたシステムの開発・運用にはデータ提供者、モデル学習者、運用者、利用者など多様な主体がステークホルダとなるが、その中で悪意あるモデル学習者が存在したとすると、危険な振る舞いの誘発やデータ漏洩など深刻な問題が生じ得る。そこで本論文では次の３つの方針をとる。1. 協調学習・フェデレーテッドラーニングのように複数のローカル学習者が関わる環境では、悪意ある学習者が勾配情報を悪用して攻撃する可能性があるが、安全な多者計算（MPC）と差分プライバシー（DP）を組み合わせ、勾配を秘匿しつつ必要なノイズ量を削減することで攻撃を防ぎながら高い精度を維持する Secure and Private Gradient Computation (SPGC)を提案する。2. 中央集権型学習では、悪意ある学習者がどこまで危険なモデルを作れるかは十分に理解されていなかったこともあり TALPA と呼ぶ新たなバックドア攻撃を提案し、入力空間・潜在空間の両方で検出されにくい強力な攻撃が可能であることを示すとともに、内部表現のクラス間異常を検知することが防御の有望な方向性であることを示す。3. 悪意ある学習者がモデルを細工することでモデル出力にユーザ入力を復元可能な情報を埋め込む攻撃を提案する。これは Trusted Execution Environment（TEE）や MPC を用いた秘匿推論環境でも成立し、入力データが秘匿されていてもプライバシー侵害されうるが、その対策として入力の多様性を高める変換が攻撃効果を低減する可能性を示す。

上記から本博士論文では、AI をベースとしたサプライチェーンにおける悪意あるモデル学習者が引き起こす攻撃と、それに対抗する防御手法を体系的に分析した手段を確立した。これにより、将来的な信頼性・安全性・プライバシーを備えた AI サプライチェーンの構築に向けた基盤を作り上げることに大きな貢献をした。よって、博士（情報科学）の学位論文として価値のあるものと認める。