STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–16.
Published online in International Academic Press (www.IAPress.org)

Framing Adversarial Federated Learning Threats through MITRE ATLAS

Tarik Guemmah [1], Hakim El Fadili [2]

[1] *Computer Science and Interdisciplinary Physics Laboratory, ENS, University Sidi Mohamed Ben Abdellah, Fez, Morocco*
[2] *Laboratory of Artificial Intelligence, Data Science and Emerging Systems, ENSAF, University Sidi Mohamed Ben Abdellah, Fez, Morocco*

Abstract As the adoption of Federated Learning (FL) accelerates across sectors prioritizing privacy, its decentralized architecture introduces novel cybersecurity threats that remain underrepresented in existing adversarial threat taxonomies. This paper bridges this gap by systematically analyzing FL-specific adversarial techniques and mapping them to the MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) framework, a living knowledge base for Artificial Intelligence Systems threats. Through a structured methodology and systematic literature review of 126 peer-reviewed articles published between 2021-2025, complemented by empirical validation through detailed case studies, we were able to find out that federated learning is vulnerable to several critical vulnerabilities, such as model poisoning, privacy leakage, and collusion attacks both in cross-silo and cross-device settings. The in-depth analysis of the current coverage in MITRE ATLAS reveals considerable weaknesses in its coverage and the mitigation measures are critically analyzed in the light of computational overhead, scalability concerns, and regulatory compliance issues. This contribution proposes extensions to the MITRE ATLAS framework, enables AI threat intelligence operationalization and provides a systematized roadmap of standardization of federated learning threat modeling in the ATLAS framework.

Keywords Adversarial Machine Learning, Federated Learning, MITRE ATLAS, Cybersecurity of Artificial Intelligence Systems.

AMS 2010 subject classifications: 68T01, 68M14, 94A60

DOI: 10.19139/soic-2310-5070-2842

1. Introduction

Artificial Intelligence (AI) is a cornerstone of new technologies and is being widely used in many fields [1]. AI has ushered in a new era of cybersecurity challenges [2] that cannot be unnoticed. In fact, as with any technology, the triad of cybersecurity related to the confidentiality, the integrity and the availability, is also applied in this context. One of many cybersecurity concerns is related to impacting the objectives and the outputs of the AI systems by introducing malicious inputs known as Adversarial Machine Learning (AML). AML attacks occur when slight human-imperceptible changes to inputs return significantly different outputs. The slight perturbations can cause significant changes in the predictions of the models or mislead the classification of the machine learning algorithms. This phenomenon reveals the vulnerability of machine learning models and opens new attack vectors. Common methods for generating adversarial samples include the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Carlini & Wagner (C&W) attack, and DeepFool [2].

In our study, we will focus on Federated Machine Learning (FL), a specific type of machine learning, as the core subject to secure. In fact, FL is a decentralized, privacy-preserving machine learning paradigm that enables

*Correspondence to: Tarik Guemmah (Email: tarik.guemmah@usmba.ac.ma). Computer Science and Interdisciplinary Physics Laboratory, ENS, University Sidi Mohamed Ben Abdellah, 5206 Bensouda Fez, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

multiple entities (data islands) to train collaboratively a shared global model without centralizing or directly exchanging raw data [4]. Instead of transferring sensitive data to a central server, FL distributes the model to local parties, where it is trained on private datasets, and only model/features updates are transmitted to a central aggregator or coordinator for synthesis into an improved global model [5]

This federated approach has emerged as a cutting-edge paradigm in artificial intelligence, addressing critical challenges in data privacy, regulatory compliance, and decentralized collaboration, making it indispensable across multiple high-stakes industries. For example, in the healthcare sector, FL enables hospitals and research institutions to collaboratively train diagnostic models, such as the prognosis of COVID-19 [6], cancer detection [6], and rare disease analysis [6], without sharing sensitive patient data. In addition, FL complies with healthcare regulations while improving diagnostic accuracy through diverse datasets. Another example related to the financial sector, FL leverages fraud detection [7] and credit risk modeling [7]. In this case, FL allows banks to refine algorithms using transaction patterns across institutions without exposing customer data, thereby enhancing security while adhering to financial privacy laws. In smart cities and using Internet of Things (IoT), FL optimizes traffic management [2], energy distribution [8], and environmental monitoring [9] by processing data locally on edge devices, reducing latency and bandwidth costs while preserving citizen privacy. The autonomous vehicle industry employs FL to improve collision detection [9] and navigation systems [9] by aggregating sensor data from distributed fleets without centralized data storage, ensuring real-time adaptability and security. Finally, retail and manufacturing benefit from FL-driven demand forecasting and supply chain optimization [9], where decentralized data from multiple locations improves predictive precision without compromising proprietary business information.

The second part of our study is related to modeling FL AML cybersecurity risks through MITRE Adversarial Threat Landscape for AI Systems (ATLAS) [10], which is a globally accessible, living knowledge base of adversary tactics and techniques based on real-world attack observations and realistic demonstrations from artificial intelligence red teams and security groups. Exploring this framework reveals a growing number of vulnerabilities in AI-enabled systems due to the increase in the attack surfaces of existing systems beyond those of traditional cyberattacks. The objective of ATLAS is to raise the awareness and readiness of the community for threats, vulnerabilities, and risks in the broader AI assurance landscape. ATLAS efforts are focused on capturing cross-community data on real-world AI incidents arising when using open-source models or data, building new open-source tools for threat emulation and AI red teaming, and developing mitigations to defend against AI security threats [10]. Our objective is to inherit and adapt the existing framework to a specific FL.

Both parts of our study lead us to focus on securing FL enabled systems and giving more attention to novel attack surface introduced by the matching of FL's decentralized nature and the threat landscape. However, even the existing MITRE ATLAS framework inadequately maps the FL risks to its matrix. In fact, the key gaps addressed in our study are :

- The lack of FL-specific Tactics, Techniques and Procedures (TTPs);
- The identification of cross-silo (enterprise-scale) vs. cross-device (edge-device) FL cyber threats.

In this study, we present, first, our review of related works related to AML, FL, and MITRE ATLAS framework, addressing a state-of-the-art and a holistic overview of the threats landscape. Then, we emphasize our contribution as, giving a methodology of mapping specific FL techniques with extensions to MITRE ATLAS tactics and applying them on a selection of Tactics related to MITRE ATLAS framework.

2. Background

2.1. Adversarial Machine Learning (AML)

To implement the slight human-imperceptible changes to inputs to Machine Learning algorithms returning significantly different outputs, many options are available. In this section, we will begin with a general representation of AML, and then we will frame our selection of methods for generating adversarial samples. In fact, as a general representation, we can address AML as follows :

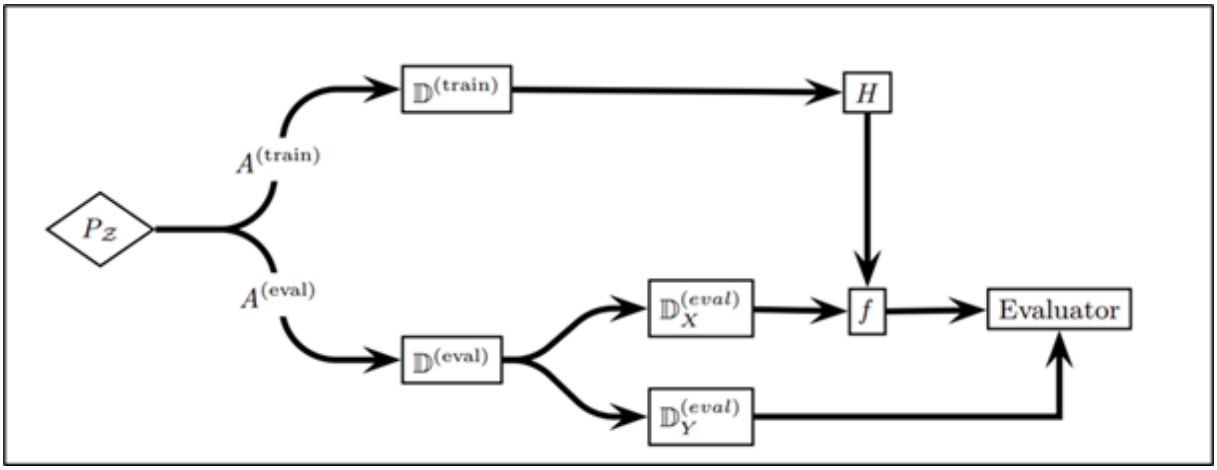

Figure 1. AML simplified representation

In Figure 1, we show a diagram of an attack against a learning system [41] where P_z is the data's true distribution, $A^{(train)}$ and $A^{(eval)}$ are adversary's attack procedures, $D^{(train)}$ and $D^{(eval)}$ are training and test datasets, H is the learning algorithm, and f is the hypothesis it learns from the training data. The hypothesis is evaluated on the test data by comparing its prediction $f(x)$ to the true label y for $(x, y) \in D^{(eval)}$.

Defining our selection of methods for generating adversarial samples, we have chosen the following :

1. The Fast Gradient Sign Method (FGSM) [11] is a foundational single-step adversarial attack that generates adversarial examples by exploiting the gradients of a model's loss function. Given a model with parameters θ, and an input-label pair (x, y), FGSM perturbs the input using the gradient of the loss function $J(\theta, x, y)$ with respect to the input, as shown in the equation:

$$x' = x + \epsilon \cdot sign(\nabla_x J(\theta, x, y)). \tag{1}$$

where, ϵ is the magnitude of the perturbation.

FGSM operates under the assumption that in high-dimensional spaces, even small perturbations can cause significant shifts in a model's decision boundaries, due to the local linearity of deep networks [12]. By following the sign of the gradient, FGSM efficiently increases the model's loss in a single step, making it computationally inexpensive. However, if the perturbations produced are large, meaning ϵ is very high, it can make them more perceptible, especially in visual data domains where stealth is essential [12, 13]. Despite this limitation, FGSM remains a fundamental benchmark in adversarial training and model robustsubness evaluation. This makes FGSM fast but suboptimal in subtlety, often resulting in adversarial examples that are easier to detect. To mitigate this, researchers have proposed refinements such as Random Start + FGSM (R+FGSM), gradient regularization, and curriculum-based adversarial training to constrain perturbation

magnitude and improve training dynamics.

2. **Projected Gradient Descent (PGD) [12]:** extends FGSM by applying iterative updates within a bounded region. At each step, the input is perturbed and projected back into an ϵ-ball surrounding the original input to ensure the attack remains within allowable distortion limits. The update rule is:

$$x^{(t+1)} = \prod_{\mathcal{B}_e(x)} \left(x^{(t)} + \alpha \cdot sign \left(\nabla_x \mathcal{L} \left(\theta, x^{(t)}, y \right) \right) \right) \tag{2}$$

where:

Symbol	Meaning
$x^{(t)}$	The adversarial example at iteration t. This is the current perturbed input.
$x^{(t+1)}$	The new adversarial example at iteration $t+1$, after applying the gradient update and projection.
α	The step size (a small scalar) that controls how big the update is at each iteration.
$\nabla_x \mathcal{L} \left(\theta, x^{(t)}, y \right)$	The gradient of the loss function $\mathcal{L}$ w.r.t. the input x, holding the model parameters θ and label y fixed. This tells us how to change the input to increase the loss.
$sign(\cdot)$	The sign function: it returns the sign (positive or negative) of each component in the gradient. This makes the update direction maximal under the ℓ_∞-norm constraint (same trick as FGSM).
$x^{(t)} + \alpha \cdot sign(\cdot)$	This is the proposed update, moving in the adversarial direction.
$\prod_{\mathcal{B}_e(x)}(\cdot)$	This is the projection operator onto the ℓ_∞-ball $\mathcal{B}_e(x)$. It ensures that the updated input stays within a bounded region (distance ϵ) around the original input x. In practice, it clips the values so that the perturbation stays within $[-\epsilon, +\epsilon]$.

3. **Carlini and Wagner (C&W) Attack [13]:** is an optimization-based method designed to generate adversarial examples with minimal distortion. The attack solves the following objective:

$$\min_{\delta} ||\delta||_p + c \cdot f(x+\delta) \tag{3}$$

where $x + \delta \in [0,1]^n$, $f(x+\delta)$ is a loss function that encourages misclassification, and c is a regularization constant.
Unlike FGSM and PGD, which rely on stepwise gradients, C&W employs gradient-based optimizers to directly minimize the perturbation while ensuring a successful attack. It is known for its ability to bypass advanced defenses like defensive distillation, although at the cost of high computational overhead. C&W is particularly useful for evaluating high-precision robustness claims.

4. **DeepFool [14]:** is an attack that iteratively linearizes the classifier decision boundary to compute the minimal perturbation necessary to change the predicted label. At each iteration, it moves the input in the direction orthogonal to the closest decision boundary. It can be represented by the following equation:

$$x^{(t+1)} = x^{(t)} - \frac{\mathcal{L}\left(x^{(t)}\right)}{||\nabla \mathcal{L}\left(x^{(t)}\right)||^2} \cdot \nabla \mathcal{L}\left(x^{(t)}\right) \tag{4}$$

where:

Symbol	Meaning
$x^{(t)}$	The current adversarial example (input) at iteration t
$x^{(t+1)}$	The updated input at iteration $t+1$ (closer to the decision boundary)
$\mathcal{L}\left(x^{(t)}\right)$	The classifier output score for the current class. It acts like a "distance" from the decision boundary
$\nabla_x \mathcal{L}\left(x^{(t)}\right)$	The gradient (direction of fastest change) of the classifier output at point $x^{(t)}$
$\|\|\nabla \mathcal{L}\left(x^{(t)}\right)\|\|^2$	The squared norm of the gradient vector (how "steep" the boundary is at this point)

DeepFool is particularly valuable for robustness analysis because it approximates the distance from a data point to the decision boundary. It is computationally efficient and effective in generating subtle adversarial examples, making it a standard method for measuring a model's vulnerability in clean (non-adversarially trained) scenarios.

Thus, each of the reviewed adversarial attack methods contributes uniquely to our understanding of model vulnerabilities. FGSM is efficient but less subtle; PGD offers a strong iterative baseline; C&W demonstrates high-precision adversarial capability; and DeepFool balances computational cost with perturbation minimality. These methods form the foundation of adversarial robustness research and are critical for evaluating and improving the security of modern AI systems. In the next section, we will discuss the cybersecurity challenges in FL-context.

2.2. Federated Learning cybersecurity challenges

Federated Machine Learning technology allows participants to build a joint training model, and maintain underlying data locally. The original concept of FL was extended to refer to all privacy-preserving decentralized collaborative machine learning techniques. FL is able to tackle not only horizontally partitioned data according to samples but also vertically partitioned data according to features in collaborative-learning setting. In Figure 2, Google has proposed a simplified framework to represent FL implementation.

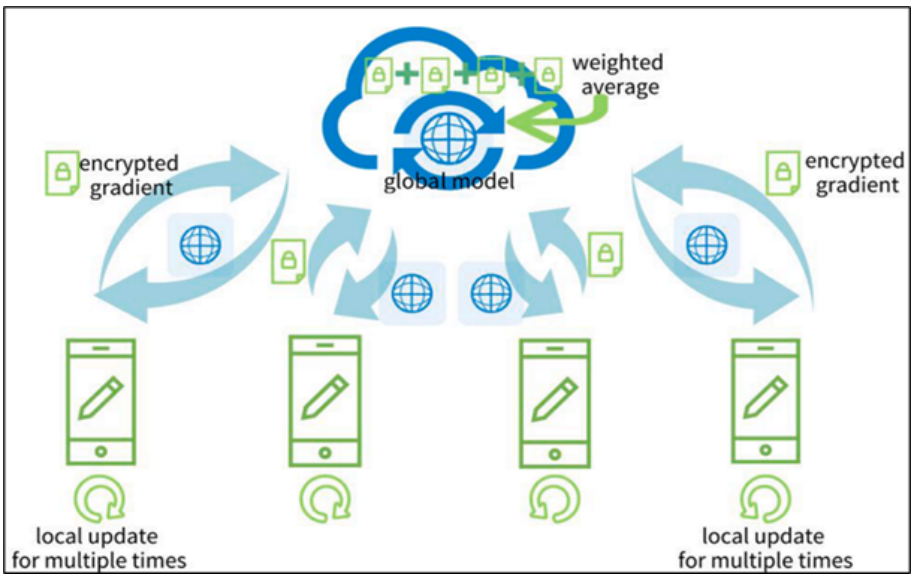

Figure 2. Illustration of FL framework proposed by Google

Federated Learning introduces a decentralized paradigm for collaborative model training while preserving data privacy [15], yet its distributed nature exposes it to unique security vulnerabilities. One critical issue is model poisoning, where malicious clients submit manipulated updates (e.g., gradient scaling, sign-flipping, or backdoor triggers) to degrade global model performance or embed hidden malicious behaviors [15]. For instance, Byzantine attacks disrupt convergence by injecting falsified updates [12], while backdoor attacks [6] subtly alter model behavior for specific inputs (e.g., misclassifying triggered medical images). Another concern is privacy leakage, where adversaries exploit gradient updates to reconstruct sensitive local data (e.g., via deep leakage from gradients (DLG) or membership inference attacks), violating FL's core privacy guarantees [12].

Additionally, free-rider attacks [14] exploit the system by submitting low-effort or duplicated updates, wasting computational resources without contributing meaningfully to model training. The lack of centralized oversight in decentralized FL exacerbates these risks, as malicious peers can propagate poisoned updates directly to others, undermining consensus. Furthermore, heterogeneity in data distributions (non-independent and identically distributed) and device capabilities complicates robust aggregation, leaving models vulnerable to biases or adversarial exploitation. While defenses like Secure Aggregation [11] (SecAgg), Differential Privacy [11] (DP), and Byzantine-robust algorithms [16] mitigate some risks, challenges persist in balancing security with model utility, scalability, and real-world deployment constraints. Emerging threats, such as cross-device collusion [17] and adaptive adversarial attacks [12], demand continued research into hybrid defenses combining cryptographic techniques (e.g., homomorphic encryption), decentralized trust frameworks (e.g., blockchain) [18], and dynamic validation mechanisms [19].

2.3. Overview of MITRE ATLAS

MITRE Adversarial Threat Landscape for AI Systems (ATLAS™) is a globally accessible, living knowledge base of adversary tactics and techniques based on real-world attack observations and realistic demonstrations from artificial intelligence (AI) red teams and security groups [20]. There are a growing number of vulnerabilities in AI-enabled systems as the incorporation of AI increases the attack surfaces of existing systems beyond those of traditional cyberattacks. ATLAS's objective is to raise community awareness and readiness for these unique threats, vulnerabilities, and risks in the broader AI assurance landscape. ATLAS community efforts are focused on capturing cross-community data on real-world AI incidents, growing understanding of vulnerabilities that can arise when using open-source models or data, building new open-source tools for threat emulation and AI red teaming, and developing mitigations to defend against AI security threats[21].

In the following, the definition of the tactics used by the adversary in MITRE ATLAS related to attacks on AI systems that were based on [33]:

1. Reconnaissance : The adversary is trying to gather information about the AI system they can use to plan future operations.
2. Resource Development : The adversary is trying to establish resources they can use to support operations.
3. Initial Access : The adversary is trying to gain access to the AI system.
4. AI Model Access : The adversary is attempting to gain some level of access to an AI model.
5. Execution : The adversary is trying to run malicious code embedded in AI artifacts or software.
6. Persistence : The adversary is trying to maintain their foothold via AI artifacts or software.
7. Privilege Escalation : The adversary is trying to gain higher-level permissions.
8. Defense Evasion : The adversary is trying to avoid being detected by AI-enabled security software.
9. Credential Access : The adversary is trying to steal account names and passwords.
10. Discovery : The adversary is trying to figure out your AI environment.
11. Collection : The adversary is trying to gather AI artifacts and other related information relevant to their goal.
12. AI Attack Staging : The adversary is leveraging their knowledge of and access to the target system to tailor the attack.
13. Command and Control : The adversary is trying to communicate with compromised AI systems to control them.

14. Exfiltration : The adversary is trying to steal AI artifacts or other information about the AI system.
15. Impact : The adversary is trying to manipulate, interrupt, erode confidence in, or destroy your AI systems and data.

In Figure 3, we show an extract of the MITRE ATLAS matrix. This framework offers a matrix representation of all Tactics, Techniques, and Procedures (TTP) related to attacks on AI systems.

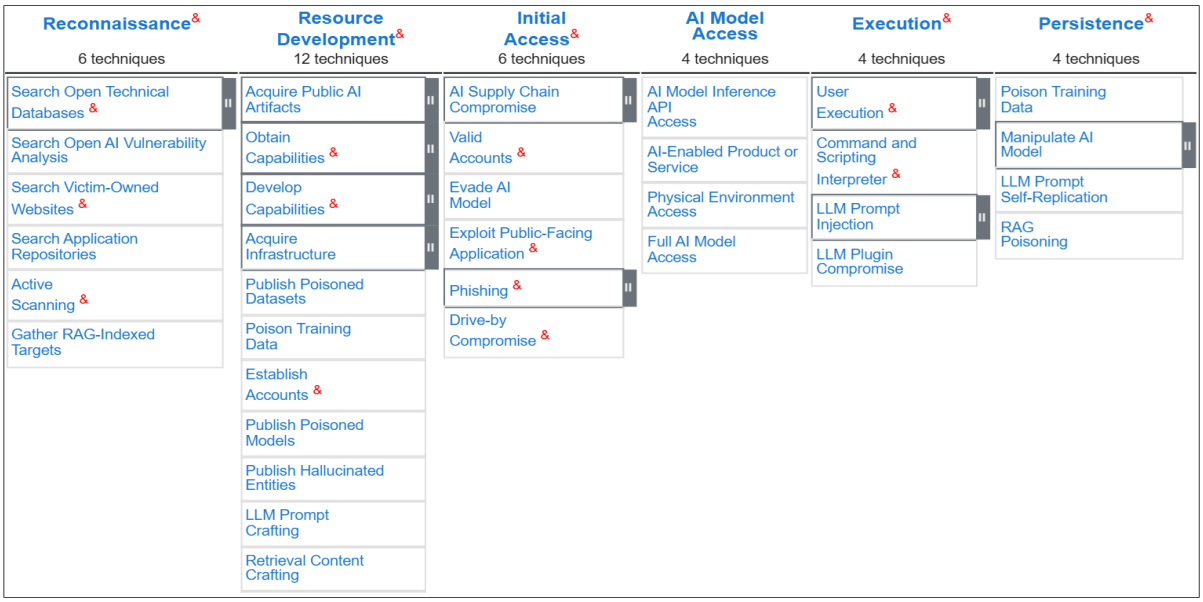

Figure 3. Extract of MITRE ATLAS framework matrix

3. Method

We suggest the following methodology for framing Adversarial Machine Learning and Federated Learning Threats through MITRE ATLAS, conducted in 4 steps, described in this section.

1. Systematic Literature Review (SLR): We searched major digital libraries (IEEE Xplore, ACM, SpringerLink) for FL security and adversarial ML publications (2021–2025)
2. ATLAS Gap Analysis: In step two, we cross-referenced identified FL attack vectors against the MITRE ATLAS v4.1 framework. Each threat was checked against existing ATLAS tactics/techniques. Unmapped behaviors were identified as gaps.
3. Technique Formulation: Step three involved proposing new ATLAS-style techniques. For each identified gap in step 2, we create an ”AML.TAxxxx” identifier consistent with the ATLAS naming. We derived the scope of each new entry by aligning the attack tactics with the ATLAS taxonomy.
4. Defense Validation: We included only mitigation from peer-reviewed studies.

3.1. Systematic Literature Review (SLR), Selection and Exclusion Criteria

To ensure a reproducible and systematic identification of FL adversarial threats, we employed a structured search query across major databases (IEEE Xplore, ACM Digital Library and SpringerLink) to identify relevant peer-reviewed studies published between January 2021 and February 2025. The core search string utilized Boolean logic, combining terms related to the system, the threat context, and the framework: (”Federated Learning” OR ”Distributed AI”) AND (”Adversarial Attack” OR ”Data Poisoning” OR ”Model Poisoning” OR ”Inference

Attack" OR "Privacy Leakage") AND ("MITRE ATLAS" OR "Threat Modeling"). Inclusion and Exclusion Criteria focused on peer-reviewed articles published from January 2021 to February 2025 whose authors put forward attack vectors or federated learning-specific mitigations. The exclusion criteria also included those studies that did not include an empirical or analytical validation element, generic AI attacks that were not implemented in decentralized federated learning, and those studies with a lower quality score under a specified quality threshold. Figure 4 details the SLR process using PRISMA-style flow diagram :

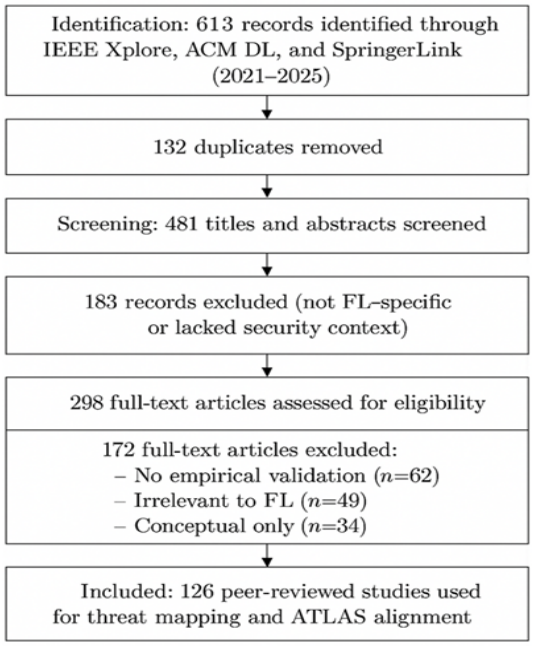

Figure 4. PRISMA-style flow diagram of study selection for systematic literature review (2021-2025).

3.2. ATLAS Gap Analysis and Technique Formulation

We used a coding-sheet-based mapping workflow so that the methodological consistency may be checked. The protocol starts with the derivation of attack properties, such as adversarial objectives and access levels they need, and then the derivation of each of the properties into an ATLAS tactic using formal definitions of adversarial intent. The next step identifies the existence of an equivalent ATLAS technique; in case none is found, the behavior is indicated as an unmapped behavior. Lastly, they are thematically clustered to produce new FL-specific techniques by using unmapped behaviors.

3.3. Quality Assessment of Selected Studies

To help counteract any bias that might have arisen during literature selection, a quality assessment rubric with a rigorously designed conceptual basis based on AMSTAR-2 was applied to every primary study [42]. Scores were assigned on a scale of 0 to 1 on five different criteria including; clarity of used threat model, methodological rigor of experiment, reproducibility of used methods, applicability to adversarial federated learning settings, and transparency in regard to the limitations that it acknowledges. Only the papers that scored (4/6) were included in the final taxonomy.

4. Results

To ensure consistency with MITRE ATLAS definitions, our mapping results are based on the adversary's operational objective rather than the attack mechanism. For example, the Federated Byzantine attack is mapped to Initial Access because the adversary's first action is to join the FL ecosystem as a malicious client, thus gaining the initial foothold required to influence model training. Another example, the Free-Rider Exploit is assigned to Resource Development because the adversary establishes computational resources (non-contributing clients) to support later phases of the attack without expending local computation. In this section, we enumerate our mapping of the proposed techniques and some practical use cases.

4.1. Mapping MITRE ATLAS proposed techniques

To illustrate our analysis, and instead of drawing up an exhaustive table of all the works, we have chosen to illustrate a sample of articles and works in Table 1 in the appendix, consolidating each ATLAS tactic with the corresponding FL-specific technique, implementation details, defenses, and representative references. In this section, we emphasize our selection of papers, relevant to the framing of AML related to FL mapped to MITRE ATLAS.

- For reconnaissance tactic, the Deep Leakage from Gradients (DLG) version 3 is used for reconstructing malicious input data from gradients [22] causing FL Parameter Leakage. As mitigation, we can use Secure Aggregation as a protocol to securely compute aggregate statistics without revealing individual data or use Differential Privacy (DP) that is a statistical technique that adds noise to protect individual data points in a dataset [31].

- For initial access, we selected Federated Byzantine as a technique using Gaussian Noise which adds random noise to gradients to ensure differential privacy [5]. Sign-flipping on the other hand is an Adversarial manipulation where signs of gradients are reversed to corrupt model updates. Mitigations are Krum Aggregation algorithm that selects updates with minimum distance to others. FoolsGold is a Defense mechanism against attacks in FL by limiting contributions from similar gradient profiles [32].

- In the persistence, we suggest to add Federated Backdoor using GAN triggers backdoors planted while using Generative Adversarial Networks [6]. As mitigation, we selected Federated Learning-based Adversarial Model Evaluation (FLAME) or SHapley Additive exPlanations (SHAP) analysis which assigns for each feature a contribution score for a particular prediction made by the model [16].

- For the collection, we chose Vertical Federated Learning (VFL) exploit using Gradient L1-norm analysis and synthetic queries. As mitigation, we cite Label Differential Privacy (LDP) and β-Variational Autoencoder (beta-VAE) for disentangling latent features in machine learning models [40].

- For resource development, we chose Free-Rider exploit using the Zero-gradien to hide updates or disrupt training and model replay. As mitigation, Shapley Values Game-theoretic approach [34] can be used to attribute model prediction to individual features. We can also use cryptographic methods to verify if a model was trained on a specific dataset.

- For FL integrity attacks as an impact tactic, we can emphasize the Class Forgetting Phenomenon (CFP) where models "forget" certain classes over time, relevant in unlearning and backdoor analysis or saliency manipulation tampering with gradient-based attribution methods[18]. As mitigation, we can use blockchain check pointing [19] to store model checkpoints securely for auditability and unlearning verification mechanisms to verify if a model has forgotten specific training data.

- For lateral movement tactic, we focus on cross silo poisoning using Internet Protocol (IP) spoofing [17] where attacker disguises source address in FL communication or using gradient averaging. As mitigation,

Decentralized Authentication Identifiers are used to secure identity mechanism in decentralized systems including FL. On the other hand, Round Trip Time (RTT) profiling is used to detect timing attacks inferring client activity or model state [40].

- For defense evasion, we selected GenAI poisoning [31] using the Wasserstein Generative Adversarial Network with Gradient Penalty and StyleGAN perturbations as targeted changes made to images in the latent space of a StyleGAN. As mitigation, we can use AI Shield including monitoring and authentication and Cross-client validation [40].

- Finally for exfiltration, we tackled the FL Intellectual Property Theft (IP Theft) using Jacobian distillation that transfers knowledge from a complex model to a simpler one by matching their input-output sensitivity (Jacobian matrices) or parameter correlation by the analysis of statistical dependencies between model parameters, often used to detect overfitting, redundancy, or hidden backdoors in neural networks [22]. As a defense mechanism, we selected dynamic watermarking used to embed and verify ownership or authenticity or QPM (Queries Per Minute) rate limiting by limiting frequency of API calls to prevent gradient or model extraction attacks [6].

4.2. Case Studies of proposed MITRE ATLAS extensions

To address the need for empirical grounding of the proposed MITRE ATLAS extensions, we introduce two case studies. These case studies illustrate how the newly proposed techniques manifest in realistic/testbed Federated Learning deployments and demonstrate how they align with ATLAS tactics.

4.2.1. Federated Backdoor via GAN Trigger Injection (AML.TA0004)

- Scenario overview: A healthcare consortium trains a federated diagnostic model on medical imaging data. One hospital becomes compromised and injects a GAN-generated trigger into its local data during training.
- Adversary steps mapped to ATLAS tactics:

  1. Initial Access (AML.TA0003): The malicious hospital joins the FL round as a legitimate participant.
  2. AI Model Access: The adversary receives the global model and can analyze its gradients.
  3. AI Attack Staging: The local data is poisoned with GAN-generated patterns designed to activate a backdoor.
  4. Execution (via update submission): The adversary sends manipulated gradients to the server.
  5. Persistence (AML.TA0004): The backdoor persists across rounds due to the aggregator's inability to distinguish poisoned updates from benign ones.
  6. Impact: The global model misclassifies triggered samples during inference.

- Outcome: The attacker achieves persistent, undetected access to influence model behavior. This validates the placement of "Federated Backdoor via Trigger Injection" under Persistence, consistent with MITRE's definition.

4.2.2. Reward Manipulation in Federated Reinforcement Learning (AML.TA0011)

- Scenario overview: A fleet of autonomous drones trains a shared navigation policy using Federated Reinforcement Learning. A compromised drone manipulates the reward signal to favor risky behaviors.
- Adversary steps mapped to ATLAS tactics:

  1. Reconnaissance: The adversary analyzes reward distributions to identify sensitive transitions.
  2. AI Model Access: The attacker obtains the policy or Q-value updates.
  3. AI Attack Staging: The local reward function is altered to mislead global convergence.
  4. Impact (AML.TA0011): The global navigation model converges toward unstable policies, reducing safety margins.

5. Persistence: Due to temporal credit assignment, the harmful bias persists across rounds without immediate detectability.

- Outcome: This validates the need for a specific ATLAS technique addressing reward manipulation and temporal exploitation unique to FRL systems.

5. Discussion

5.1. Underexplored Threat Vectors in MITRE ATLAS

5.1.1. ***Federated Reinforcement Learning (FRL)*** is a subset of FL where several agents train reinforcement learning models without sharing environmental events or reward signals [22]. FRL extends FL's decentralized and privacy-preserving principles to sequential decision-making and complex real-time systems like autonomous cars, drone swarms, robotic control, and smart grid coordination. Autonomous car fleets learn their navigation policies locally and update a worldwide driving policy without exposing trip data [15]. While FRL inherits the privacy benefits of FL, it introduces a set of novel and underexplored vulnerabilities due to the delayed, cumulative, and environment-dependent nature of reinforcement feedback. Reward poisoning, when a malevolent actor quietly manipulates the reward signal to favor poor or risky global policy behavior, is a major vector. In a self-driving vehicle, a malevolent client may report that braking late at intersections rewards more, biasing the shared model toward dangerous behavior. This manipulation is harmful because it typically goes unnoticed in early training rounds, appearing only after convergence when model behaviors are strongly established in system policy [22].

Another risk is policy drift via delayed manipulation. FRL attacks may wait several rounds before diverging the policy space, unlike supervised learning attacks that misclassify or degrade models immediately. This renders anomaly detection methods useless. The absence of global insight into local settings allows attackers to add environment-specific biases that generalize poorly and harm global performance. Furthermore, multi-agent FRL systems, such as collaborative robotic arms or drone fleets, introduce the risk of collusion-based attacks. A subset of adversarial agents may synchronize gradient updates or imitate experiences to influence global policy. Parameter aggregation connects conceptually separated agents, making inter-agent signaling exploitation hard to detect. Without precise coordination modeling, the system cannot differentiate innocuous updates from organized manipulation [23]. Despite these risks, MITRE ATLAS lacks a sub-technique or strategy to classify these assaults. In the "Impact" phase, we propose AML.TA0011: Reward Manipulation and Temporal Policy Exploitation in FRL Systems. This method would catch:

- **Reward Bias Injection:** tampering with local scalar feedback to misguide global learning.
- **Policy Drift:** delayed-impact manipulation leading to suboptimal convergence.
- **Collusive Agent Behavior:** synchronized adversarial updates in multi-agent settings.
- **Environment-overfitting:** Agents sharing gradients optimized for local environments that generalize poorly.

This method will address a major ATLAS matrix gap and allow security analysts and system designers to analyze and manage FRL-specific hazards. Examples from real life demonstrate the importance of this expansion. Waymo and Tesla use fleet-wide decentralized learning algorithms for autonomous navigation. One hacked car might influence lane-keeping algorithms by reporting false incentives for GPS errors or dangerous moves. Even a small number of antagonistic agents may mislead coordinated drones used for agriculture or defense into suboptimal formations or coverage gaps in swarm robotics. These attack methods differ from adversarial machine learning (AML), which targets instantaneous inference outputs. Instead, FRL assaults use temporal dynamics, policy generalization, and the central aggregator's incapacity to audit experiences. ATLAS will help practitioners create security measures including policy regularization, adversarial agent filtering, and reward auditing by formalizing these assaults. Federated Reinforcement Learning creates new attack surfaces that conventional adversarial threat frameworks cannot manage. AML.TA0011 is a specialized technique that recognizes the rising relevance of timed, delayed, and coordinated assaults in decentralized systems and forms the basis for proactive defensive development.

5.1.2. ***Cross-Silo vs. Cross-Device FL Threat Surfaces*** : Current ATLAS mappings fail to distinguish cross-silo and cross-device FL deployments, which is another major restriction. This mistake hinders ATLAS's ability to represent real-world federated systems, where deployment design greatly affects threat characteristics. A few well-established organizations including hospitals, banks, and universities collaborate in cross-silo FL. These organizations usually have reliable infrastructures, skilled IT workers, and clear regulations. Structured data and dependable communication are the typical challenges. This setting often involves Byzantine assaults, model inversion via insider leaks, or targeted poisoning to impact certain activities (e.g., biased credit rating). In contrast, cross-device FL involves enormously dispersed networks with hundreds to millions of edge devices like smartphones, wearables, and IoT sensors. These devices have inconsistent connection, limited processing capabilities, and uneven data distributions. Free-rider attacks, gradient obfuscation, device ID spoofing, and compromised edge node cooperation are threats against those devices. In addition, malicious updates are more likely due to low device security. These parameters vary in size, topology, and trust assumptions, affecting attack capability, impact, and detectability. Krum aggregation [24] resists outliers in small cross-silo installations, but scalability difficulties and noisy updates from weak devices make it inefficient or useless in cross-device FL.

However, current MITRE ATLAS methods do not discriminate between deployment types. Most FL mappings assume a homogeneous threat model, overgeneralizing or underspecifying. System architects trying to threat model FL situations like hospital medical imaging vs cell phone keyboard prediction face this issue. To remedy this, MITRE ATLAS should explicitly identify technique items with deployment context tags like:

- [CS] Cross-Silo Compatible
- [CD] Compatible across devices
- [ME] Mixed Environment

This context-aware categorization will improve ATLAS entry interpretation and let stakeholders select mitigations depending on deployment factors. ATLAS should also include edge-specific tactics like device farm simulation, where adversaries simulate hundreds of virtual clients to bias federated training without physical devices, in its "Reconnaissance" and "Resource Development" tactics. Many cross-device FL systems use non-independent and Identically Distributed data regimes, where each device has its own data distribution. Aggregation techniques must be resistant against enemies and flexible to distributional skew, complicating defensive design. The absence of ATLAS representation for these statistical issues highlights the need for stronger adversarial threat model-federated system behavior alignment.

5.2. Mitigation Challenges

5.2.1. ***Privacy-Utility Tradeoffs Differential Privacy (DP)*** is one of the most widely adopted techniques for protecting user data in machine learning, including FL systems[25, 26, 27]. It hides data point contributions by adding calibrated noise into model updates or gradients. The key is to guarantee that exclusion or inclusion of a single data sample does not substantially impact function output, as measured by privacy loss parameter ϵ (epsilon). Lower ϵ-values suggest better privacy but greater noise levels [28]. In federated situations, DP is usually achieved via local differential privacy (LDP), where noise is introduced to client updates before transmission, or global differential privacy, where noise is added to the server-side aggregated model. Both techniques hinder model convergence, generalization, and usefulness. Differential Privacy (DP) is an appropriate solution to protect the data, but it faces significant scalability issues when applied to cross-editing learning with millions of users. The scale of noise needed to support the maintenance of a strong privacy guarantee often triggers significant utility loss, especially when the data is sparse, e.g., in diagnostics of rare diseases. In addition, the communication overhead of DP combined with Secure Aggregation makes the technology worse in terms of latency and battery consumption, which makes this method frequently inapplicable to mobile edge devices. Mission-critical areas seldom allow this performance decrease.

Additionally, DP affects tasks differently. In balanced classification tasks like digit recognition, performance reductions may be acceptable. However, in unbalanced and sparse datasets like fraud detection and rare illness

diagnosis, injected noise disproportionately impacts underrepresented groups, exacerbating model biases [30]. DP increases delay and communication inefficiency. Noisy updates converge slower and need more communication rounds, particularly with Secure Aggregation or Homomorphic Encryption. This may increase device battery consumption, bandwidth utilization, and update delays in real scenarios like mobile phone federated learning, reducing FL framework scalability and responsiveness [31].

DP assumes the private function's sensitivity is known. Due to data heterogeneity (non-IID distributions), sensitivity assessment is challenging and leads to inconsistently calibrated noise in FL. The noise distribution and clipping standards determine whether this miscalibration reduces privacy or model performance [31].
In general, DP ensures privacy but is difficult to apply in FL systems. Dynamic calibration, adaptive clipping, and hybrid solutions that integrate DP with other security procedures such as aggregation masking or tailored models are needed to balance privacy and usefulness.

5.2.2. ***Free-Rider Exploits and Adaptive Adversaries*** are FL mitigation's second biggest issue. Adaptive adversaries monitor global model changes, find aggregation logic holes, and conduct covert, delayed, or coordinated assaults to escape detection [32]. In model poisoning assaults, an adaptive adversary may appear benignly to earn confidence, then progressively increase its gradient manipulation. This sluggish solution avoids classical anomaly detection methods like norm clipping and update filtering, which look for substantial departures from the average. Backdoor persistence occurs when an attacker embeds a trigger in the model (e.g., a pixel pattern in an image or a keyphrase in text) and quietly alters it during training. These shifting triggers are tougher to identify and eliminate using unlearning or reweighting, particularly when the attacker replicates benign client behavior [33]. A similar but separate issue is free-rider assaults. Training is dispersed among several clients in FL, and model aggregation requires voluntary involvement and local computation. Malicious or sluggish clients provide untrained (random or outdated) model changes or replicate other clients' updates without learning. Clients abuse the mechanism to get the final global model without investing [34].

Free-riders' updates might be statistically comparable to innocuous low-resource devices, making detection challenging. Update entropy and gradient variance fail in diverse situations. Proof-of-Learning (PoL) and Model auditing is a normative standard based on Shapley values. However, these values are too computational to be scaled in large federated learning systems. In this regard, we suggest introducing Approximate Shapley Values or gradient-variance profiling, which are also relatively complex to compute but provide sufficient auditability to real-world applications [35]. Adaptive opponents that disguise their poisoned updates as under-trained or resource-limited contributions make these behaviors more deadly. Defense co-evolution issues arise when new defenses spur the development of more complex assault techniques, creating a cat-and-mouse game. To fight against adaptable and freeloading customers, defensive measures must adjust. Machine unlearning, temporal client behavior profiling, and cryptographic local training verification may be used. Such techniques are still developing and must be tuned carefully to avoid punishing genuine players in noisy or limited contexts.

To illustrate how an adaptive adversary could evade existing defenses, consider a federated healthcare application where hospitals collaboratively train a diagnostic model. An adaptive attacker initially behaves like a benign client, sending gradient updates consistent with legitimate learning to build trust. Over successive rounds, the attacker slowly injects small but targeted backdoor patterns into its updates, remaining within the thresholds of anomaly detection mechanisms. By gradually amplifying the perturbation and periodically mimicking low-resource (free-rider) behavior, the adversary evades standard defenses designed to flag abrupt or large deviations.

5.2.3. ***Regulatory Misalignment*** : FL is widely advertised as a privacy-preserving technology compliant with contemporary data protection rules, yet legal expectations and FL system behavior are diverging. Auditability and cross-jurisdictional compliance are major concerns. Article 5 of the GDPR establishes transparency, data minimization, and accountability. Companies must explain algorithmic judgments and keep data processing logs. FL complicates these commitments in various ways. Since raw data never leaves the client device, it's hard to

tell how training data affects the model [36]. Even worse, privacy-preserving methods like Secure Aggregation, DP, and Homomorphic Encryption actively hide model update origins. This improves user privacy but hinders interpretability and explainability, particularly in regulated industries like healthcare, banking, and criminal justice. FL introduces an auditability paradox: technologies that safeguard users' privacy prohibit meaningful system auditing. HIPAA presents comparable issues. FL enables decentralized model training across hospitals or clinics without centralizing patient data, supposedly complying with data-sharing regulations. However, model malfunctions or privacy breaches (e.g., membership inference) make forensic investigation and legal liability problematic. Who owns the model? The hospital upgrade that caused the vulnerability? Current laws do not address these issues [37].

Additionally, cross-border FL partnerships create jurisdictional issues. GDPR, HIPAA, and local data localization rules must be followed concurrently by federated hospitals in the EU, USA, and Southeast Asia. FL systems generally lack centralized control or data custodians, making multi-jurisdictional compliance difficult administratively and legally. Algorithmic fairness auditing remains unsolved. Regulatory advice increasingly mandates AI systems to be racially, gender, and socioeconomically neutral. Federated systems make fairness evaluation harder, particularly when protected characteristics are not shared and client data is non-IID. This may lead to unnoticed and uncorrected biases. To overcome the auditability paradox in the regulation systems of both the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA), we have suggested a series of technical mechanisms. To start with, the scheme of Verifiable Update Tagging allows the server to confirm the integrity of an update, without any knowledge of the underlying information. Second, the use of Zero-Knowledge Proofs (ZKPs) allows the clients to prove that they were following the recommended training specifications without revealing sensitive local parameters. Federated accountability logs (tracking all modelupdates, participating clients, and aggregation steps), verifiable update tagging (origins integrity checks), and zero-knowledge proofs, without revealing any underlying information may prove compliance without disclosing sensitive information [37].

6. Conclusion

The paper is a step further than a conceptual review because it suggests a hierarchical plan of incorporating federated learning threats into the MITRE ATLAS framework. It provides a roadmap of context-sensitive tagging and empirical case studies to frame a potential foundation of practitioners to operationalize AI threat intelligence in high-stakes federated settings. In a systematic literature review of over 120 peer-reviewed studies published between 2021 and 2025, we found critical FL vulnerabilities like model poisoning, privacy leakage, free-rider exploitation, and reinforcement learning manipulation that are underrepresented or absent in current threat taxonomies. A fundamental contribution of this study is the planned expansion of MITRE ATLAS to include FL-specific attack methodologies and tactics. Our mapping identified significant ATLAS coverage gaps, specifically in reward poisoning in federated reinforcement learning, cross-device collusion, and privacy-utility tradeoff issues. Based on proven empirical data, we presented new method IDs (e.g., AML.T0107–AML.T0112) and documented their technological implementations and defenses to fill these gaps.

We highlighted mitigating problems that limit safe FL system implementation in real life. The emergence of adaptive adversaries, differential privacy, and the impossibility of present FL architectures to achieve GDPR auditability and HIPAA transparency degrade model usefulness. Structured defensive orchestration is needed because privacy, scalability, and robustness tradeoffs exist. We suggested three steps to address these limitations: (1) creating an ATLAS-FL Working Group to formalize threat mappings and develop red-teaming toolkits; (2) integrating certified robustness techniques like Zikertort proofs and Proof-of-Learning; and (3) designing a Federated Threat Mitigation Framework (FTMF) to adapt defenses to specific contexts like healthcare or IoT. This is a major step toward operationalizing adversarial intelligence for decentralized AI. The proposed ATLAS

additions and realistic roadmap provide a scalable, transparent, and community-driven mechanism to strengthen FL systems against changing adversarial threats.

Appendix

Table 1. Our consolidated mapping of AML to FL

MITRE ATLAS Tactic	Proposed Technique (AML.TXXXX)	Technical Implementation	Defense Mechanisms	Key References
Reconnaissance (AML.TA0002)	FL Parameter Leakage	• DLGv3: 50+ gradient queries • BatchNorm inversion	• Secure Aggregation • Differential Privacy • Gradient quantization	[22], [31]
Initial Access (AML.TA0003)	Federated Byzantine	• Gaussian noise • Sign-flipping	• Krum aggregation • FoolsGold	[5], [32]
Persistence (AML.TA0004)	Federated Backdoor	• GAN triggers • NLP context triggers	• FLAME • SHAP analysis	[6], [16]
Collection (AML.TA0005)	VFL Exploit	• Gradient L1-norm analysis • Synthetic queries	• Label DP • Feature β-VAE	[1], [40]
Resource Dev. (AML.TA0006)	Free-Rider Exploit	• Zero-gradient • Model replay	• Shapley values • Proof-of-Learning	[34], [8]
Impact (AML.TA0007)	FL Integrity Attack	• Class forgetting • Saliency manipulation	• Blockchain checkpointing • Unlearning verification	[18], [19]
Lateral Movement (AML.TA0008)	Cross-Silo Poisoning	• IP spoofing • Gradient averaging	• Decentralized Auth. • RTT profiling	[17], [40]
Defense Evasion (AML.TA0009)	GenAI Poisoning	• WGAN-GP EHRs • StyleGAN perturbations	• AI Shield • Cross-client validation	[31], [40]
Exfiltration (AML.TA0010)	FL IP Theft	• Jacobian distillation • Parameter correlation	• Dynamic watermarking • QPM limiting	[22], [6]
Impact (AML.TA0011)	Reward Manipulation / Temporal Exploitation	• Reward bias injection • Policy drift	• Collusive agents • Env. overfitting	[22], [23]

REFERENCES

1. Mulder, V., Humbert, M. (2023). "Differential Privacy". In: Mulder, V., Mermoud, A., Lenders, V., Tellenbach, B. (eds) *Trends in Data Protection and Encryption Technologies*, Springer
2. R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," in Proc. *22nd ACM SIGSAC Conference on Computer and Communications Security* (CCS '15), pp. 1310–1321, 2015. doi: 10.1145/2810103.2813687
3. N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data," in Proc. *5th International Conference on Learning Representations (ICLR)*, 2017.
4. Wu, X., Huang, F., Hu, Z. and Huang, H. 2023. "Faster Adaptive Federated Learning". Proceedings of the *AAAI Conference on Artificial Intelligence*, vol. 37, 9 (Jun. 2023), 10379-10387.
5. X. Fang, Z. Liu, and M. Liu, "Local Model Poisoning Attacks to Byzantine-Robust Federated Learning," in Proc. *29th ACM Conference on Computer and Communications Security* (CCS '22), pp. 1391–1406, 2022. doi: 10.1145/3548606.3560663
6. X. Shen, S. Yu, and L. Zhang, "Backdoor Attacks and Defenses in Federated Learning: A Survey," *IEEE Internet of Things Journal*, vol. 10, no. 2, pp. 857–874, Jan. 2023. doi: 10.1109/JIOT.2022.3210451

7. J. Kang, Z. Xiong, D. Niyato, D. Ye, and Y. Zhang, "Incentive Design for Efficient Federated Learning in Mobile Networks: A Contract Theory Approach," *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 6785–6799, Aug. 2022. doi: 10.1109/TWC.2022.3158039
8. H. Jia et al., "Proof-of-Learning: Definitions and Practice," 2021 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2021, pp. 1039-1056, doi: 10.1109/SP40001.2021.00106.
9. European Data Protection Supervisor. and Agencia Española de Protección de Datos, AEPD., TechDispatch: federated learning. LU: Publications Office, 2025. Accessed: Aug. 05, 2025. [Online]. Available: https://data.europa.eu/doi/10.2804/5357101
10. U.S. Department of Health & Human Services (HHS), "HIPAA and Health IT," HealthIT.gov, 2022. [Online]. Accessed: July 20, 2025. [Online]. Available: https://www.healthit.gov/topic/privacy-security-and-hipaa
11. F. Hu, W. Zhou, K. Liao, H. Li and D. Tong, "Toward Federated Learning Models Resistant to Adversarial Attacks," in *IEEE Internet of Things Journal*, vol. 10, no. 19, pp. 16917-16930, 2023.
12. R. Yu, M. R. Hesamzadeh, and D. Niyato, "Proof-of-Learning: Making Training Verifiable," in Proc. *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=9gQNnB4GZc
13. Chen, E., Cao, Y. and Ge, Y. 2024. "A Generalized Shuffle Framework for Privacy Amplification: Strengthening Privacy Guarantees and Enhancing Utility", in *AAAI Conference on Artificial Intelligence*, 38, 10 (Mar. 2024)
14. Wen, J., Zhang, Z., Lan, Y. et al. "A survey on federated learning: challenges and applications", in *Int. J. Mach. Learn. & Cyber*, 14, 513–535 (2023).
15. Y. Wei et al., "Distributed Differential Privacy via Shuffling Versus Aggregation: A Curious Study," in *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2501-2516, 2024.
16. C. Liu, J. Liu, J. Zhang, and Y. Zhang, "On Certified Robustness Against Backdoor Attacks for Federated Learning," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 3, pp. 1740–1754, May 2023. doi: 10.1109/TDSC.2022.3179647
17. A. Ghosh, N. Suri, and M. B. Amin, "FLEMON: Federated Learning and Monitoring Framework for Trustworthy Edge AI," *IEEE Internet of Things Journal*, vol. 10, no. 2, pp. 1303–1317, Jan. 2023. doi: 10.1109/JIOT.2022.3187523
18. Arun Sekar Rajasekaran, Maria Azees, Fadi Al-Turjman, "A comprehensive survey on blockchain technology", *Sustainable Energy Technologies and Assessments*, Volume 52, Part A, 2022.
19. H. Duan, Z. Peng, L. Xiang, Y. Hu and B. Li, "A Verifiable and Privacy-Preserving Federated Learning Training Framework," in *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 5, pp. 5046-5058, Sept.-Oct. 2024.
20. Liu, P., Xu, X. & Wang, W. "Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives". Cybersecurity 5, 4, 2022.
21. H. Fereidooni et al., "SAFELearn: Secure Aggregation for private FEderated Learning," in *2021 IEEE Security and Privacy Workshops (SPW)*, San Francisco, CA, USA, 2021.
22. J. Geng et al., "Improved Gradient Inversion Attacks and Defenses in Federated Learning," in *IEEE Transactions on Big Data*, vol. 10, no. 6, pp. 839-850, Dec. 2024
23. European Data Protection Supervisor (EDPS), "TechDispatch on Federated Learning," EDPS Publications, 2025. [Online]. Available: https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2025-06-10-techdispatch-12025-federated-learning_en
24. B. Han et al., "Dynamic Incentive Design for Federated Learning Based on Consortium Blockchain Using a Stackelberg Game," in *IEEE Access*, vol. 12, pp. 160267-160283, 2024.
25. Y. Jin, H. Zhu, J. Xu, and Y. Chen, *Federated Learning: Fundamentals and Advances, Singapore*: Springer, 2023.
26. K. Pillutla, S. M. Kakade and Z. Harchaoui, "Robust Aggregation for Federated Learning," in *IEEE Transactions on Signal Processing*, vol. 70, pp. 1142-1154, 2022.
27. Feng, Y., Guo, Y., Hou, Y. et al. "A survey of security threats in federated learning", in *Complex Intell. Syst.*, 11, 165, 2025.
28. G. K. J. Hussain and G. Manoj, "Federated Learning: A Survey of a New Approach to Machine Learning," in *First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)* Trichy, India, 2022.
29. Mittone, G., Riviera, W., Colonnelli, I., Birke, R., Aldinucci, "Model-Agnostic Federated Learning" In: Cano, J., Dikaiakos, M.D., Papadopoulos, G.A., Pericàs, M., Sakellariou, R. (eds) Euro-Par Parallel Processing. Euro-Par 2023. Lecture Notes in Computer Science, vol 14100. Springer, 2023.
30. P. Zhou, Q. Lin, D. Loghin, B. C. Ooi, Y. Wu and H. Yu, "Communication-efficient Decentralized Machine Learning over Heterogeneous Networks," in *IEEE 37th International Conference on Data Engineering (ICDE)*, Chania, Greece, 2021.
31. Nuria Rodríguez-Barroso, Daniel Jiménez-López, M. Victoria Luzón, Francisco Herrera, Eugenio Martínez-Cámara, "Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges", in *Information Fusion*, Volume 90, 2023.
32. M. Fang et al., "Byzantine-Robust Decentralized Federated Learning," in Proc. *ACM SIGSAC Conference on Computer and Communications Security* (CCS '24), pp. 2874–2888, 2024.
33. MITRE, "Adversarial Threat Landscape for AI Systems (ATLAS)," MITRE ATLAS Portal. Accessed: Jun. 28, 2025. [Online]. Available: https://atlas.mitre.org/
34. Run Yang, Hui He, Yulong Wang, Yue Qu, Weizhe Zhang, "Dependable federated learning for IoT intrusion detection against poisoning attacks", in *Computers & Security*, Volume 132, 2023.
35. National Institute of Standards and Technology (NIST), "AI Risk Management Framework," NIST, 2023. doi: 10.6028/NIST.AI.100-1
36. E. C. Pinto Neto, S. Sadeghi, X. Zhang, and S. Dadkhah, "Federated Reinforcement Learning in IoT: Applications, Opportunities and Open Challenges," in *Appl. Sci.*, vol. 13, no. 11, p. 6497, May 2023.
37. R. Firouzi, R. Rahmani, and T. Kanter, "Federated Learning for Distributed Reasoning on Edge Computing," *Procedia Comput. Sci.*, vol. 184, pp. 419–427, 2021.
38. Warnat-Herresthal, S., Schultze, H., Shastry, K.L. et al., "Swarm Learning for decentralized and confidential clinical machine learning", *Nature*, 594, 265–270, 2021.

39. W. Zhou, D. Zhang, H. Wang, J. Li and M. Jiang, "A Meta-Reinforcement Learning-Based Poisoning Attack Framework Against Federated Learning," in *IEEE Access*, vol. 13, pp. 28628-28644, 2025.
40. Hu, K., Gong, S., Zhang, Q. et al., "An overview of implementing security and privacy in federated learning", in *Artif Intell* Rev 57, 204, 2024.
41. Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J. D. Tygar, "Adversarial machine learning" in *4th ACM workshop on Security and artificial intelligence (AISec '11)*, Association for Computing Machinery, New York, NY, USA, 43–58, 2011.
42. B. J. Shea et al., "AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both," BMJ, p. j4008, Sep. 2017. Accessed: Jan. 1, 2026. [Online]. Available: https://doi.org/10.1136/bmj.j4008