© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723116

Securing Deep Learning Systems: Attacks and Defenses across Neural Networks, Federated, Transfer, and Reinforcement Learning

MAMOON AHMED YAHAY AL KHADHER
Master of Science in Information Technology (Data Science), School of Science & Computer Studies, CMR University Bengaluru

Abstract- ***Deep Learning (DL) techniques are widely deployed in critical applications such as autonomous driving, healthcare, and intelligent infrastructure. Core paradigms—Deep Neural Networks (DNNs), Deep Reinforcement Learning (DRL), Federated Learning (FL), and Transfer Learning (TL)—remain vulnerable to adversarial attacks that can degrade performance, leak private data, or produce unsafe decisions. Developing effective attacks and corresponding countermeasures is a prerequisite for robust, secure, and deployable artificial intelligence. Prior surveys often focused on only one or two techniques, omitted detailed discussion of datasets, metrics, and testbeds, or became outdated. This survey comprehensively reviews attacks and defenses across DNN, DRL, FL, and TL. We summarize threat models, representative attack and defense techniques, evaluation metrics, commonly used datasets, and experimental settings. A key contribution is an explicit analysis of the commonalities and differences among the four paradigms. Insights, lessons learned, and future research directions are presented to guide the development of trustworthy deep-learning systems.***

Index Terms— ***Adversarial attacks, defenses, deep neural networks, federated learning, transfer learning, deep reinforcement learning, robustness, privacy, trustworthy AI.***

I. INTRODUCTION

A. Motivation

Over the past decade, deep learning has achieved remarkable success across computer vision, natural language processing, robotics, and autonomous systems. At the same time, the discovery of adversarial examples revealed that high-performing models can be highly fragile: small, carefully crafted perturbations to inputs can cause large changes in outputs while remaining nearly imperceptible to humans. This fragility has motivated two closely related research directions—adversarial deep learning (the study of attacks) and robust deep learning (the study of defenses).

Early work concentrated primarily on standard Deep Neural Networks (DNNs) for classification. As more advanced paradigms matured, new attack surfaces appeared. Federated Learning (FL) was introduced to train models collaboratively without centralizing raw data, yet it remains vulnerable to poisoning and inference attacks. Transfer Learning (TL) accelerates training by reusing pretrained knowledge but can inherit and amplify backdoors present in source models. Deep Reinforcement Learning (DRL) underpins sequential decision-making in safety-critical domains; adversarial perturbations to observations or rewards can produce policies that endanger human safety.

Consequently, a comprehensive understanding of attacks and countermeasures across these four families is essential for building systems that are not only accurate but also robust, private, and deployable.

B. Scope and Contributions

This survey provides an integrated treatment of adversarial attacks and defenses for DNNs, FL, TL, and DRL. The main contributions are:

1. A structured review of threat models, representative attacks, and defense techniques for each of the four paradigms.
2. A summary of evaluation metrics, widely used datasets, and experimental testbeds that support reproducible research.
3. An explicit comparison of commonalities and differences among DNN, FL, TL, and DRL attack and defense landscapes.

4. A synthesis of practical insights, limitations of current approaches, and prioritized future research directions.

C. Comparison with Existing Surveys

Previous surveys typically examined only one or two techniques. DNN-focused surveys often limited discussion to digital image classification and under-emphasized physical attacks and certified defenses. FL surveys frequently concentrated on horizontal FL or a narrow subset of poisoning and inference attacks. DRL and TL surveys remained comparatively sparse and rarely placed the techniques inside a unified threat-model framework. Few works systematically catalogued datasets, metrics, and testbeds. The present survey addresses these gaps by offering a cross-paradigm perspective and by emphasizing practical evaluation resources.

II. ADVERSARIAL ATTACKS AND DEFENSES IN DEEP NEURAL NETWORKS

A. Background in Deep Neural Networks

Deep neural networks consist of multiple layers of interconnected units that learn hierarchical representations from data. Training typically minimizes a loss function—most commonly cross-entropy for classification—via stochastic gradient descent or its variants. At inference the trained network maps new inputs to predictions. Because the decision surfaces learned by modern DNNs are high-dimensional and highly non-linear, the networks can exhibit high accuracy on clean data while remaining sensitive to small input changes.

This sensitivity is the root of adversarial vulnerability. An adversarial example is an input modified by a perturbation of bounded magnitude so that the model produces an incorrect output. When the perturbation is constrained to be imperceptible (for example under an $\ell\infty$ or ℓ_2 norm bound), the attack becomes particularly concerning for security- and safety-critical applications.

B. Threat Models and Attack Taxonomy for DNNs

Attacks are commonly characterized along several axes:

• Knowledge: White-box attacks assume full access to architecture, parameters, and gradients; black-box attacks rely only on input–output queries or transfer from a surrogate model.
• Goal: Targeted attacks force a specific wrong class; untargeted attacks merely induce any misclassification.
• Phase: Training-time attacks (poisoning, backdoors) manipulate the data or optimization process; inference-time attacks craft adversarial examples against a fixed model.
• Domain: Digital attacks operate on numerical inputs; physical attacks must survive real-world sensing conditions.

C. Representative Attacks on DNNs

Gradient-based white-box attacks such as the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), the Carlini–Wagner (C&W) attack, and DeepFool remain foundational. Iterative methods (PGD and variants) generally achieve higher success rates than single-step methods under the same perturbation budget. Optimization-based attacks (C&W) produce very small perturbations at higher computational cost.

Black-box attacks frequently exploit transferability: adversarial examples crafted on a surrogate model often fool a different target model. Query-based attacks estimate gradients or search the input space directly. Backdoor (Trojan) attacks poison a fraction of the training set with a trigger pattern so that the model produces an attacker-chosen label whenever the trigger appears. Physical-world attacks such as adversarial patches demonstrate that digital vulnerability can translate into practical threats against deployed vision systems.

D. Defenses for DNNs

Adversarial training injects adversarial examples into the training loop, teaching the model to resist a chosen attack. It is computationally expensive and robustness is typically limited to the training threat model. Certified defenses provide provable guarantees: randomized smoothing certifies robustness within an ℓ_2 ball; interval bound propagation and linear relaxation methods propagate bounds through the network for a given input.

Detection methods identify adversarial inputs at test time using statistical tests, auxiliary detectors, or consistency checks. Input-preprocessing defenses (denoising, compression, feature squeezing) can remove some perturbations but are frequently circumvented by adaptive attackers. Ensemble methods and defensive distillation offer limited additional robustness against strong adaptive adversaries.

E. Insights and Limitations

Most published results focus on image classification benchmarks (MNIST, CIFAR, ImageNet). Robustness numbers often fail to transfer across architectures, datasets, or threat models. The computational cost of strong adversarial training limits adoption for the largest models. Physical realizability, multi-modal data, and sequential settings remain less explored. Rigorous evaluation under fully adaptive adversaries is not yet universal, which can produce over-optimistic robustness claims.

III. ADVERSARIAL ATTACKS AND DEFENSES IN FEDERATED LEARNING

A. Background in Federated Learning

Federated Learning enables multiple clients to collaboratively train a shared model while keeping raw data on the client devices. In Horizontal FL (HFL) clients share the same feature space but hold different samples; in Vertical FL (VFL) clients share sample identifiers but possess different feature subsets. A central server aggregates client updates, most commonly by Federated Averaging (FedAvg) or a robust variant. FL improves privacy relative to centralized training, yet the aggregation interface and the possible presence of malicious clients create distinctive attack surfaces.

B. Attacks Unique to Federated Learning

Data-poisoning attacks inject corrupted samples so that local updates push the global model toward an attacker objective. Model-poisoning attacks craft the updates themselves, often with greater impact because the attacker directly controls the message sent to the server. Backdoor attacks implant a trigger that activates only on specific inputs; because the trigger may appear rarely, validation accuracy remains high and the attack is hard to detect.

Inference attacks exploit information in shared updates or intermediate representations. Membership inference determines whether a record was in a client's training set. Property inference recovers global statistical properties of private data. Model inversion and gradient-inversion techniques attempt to reconstruct training examples from gradients or model snapshots. Byzantine attacks send arbitrary or damaging updates to prevent convergence. Colluding minorities of clients can amplify many of these attacks.

C. Defenses in Federated Learning

Robust aggregation rules reduce the influence of outlier updates. Krum and multi-Krum select updates closest to a majority; coordinate-wise median and trimmed mean limit extreme values; geometric-median-based rules provide further resilience. Differential privacy adds calibrated noise to updates or the aggregated model, providing formal privacy guarantees at a cost in utility.

Secure aggregation protocols based on secure multi-party computation prevent the server from inspecting individual updates, mitigating certain inference attacks. Anomaly detection, client reputation systems, and contribution scoring help identify suspicious participants. In practice, combinations of robust aggregation, privacy mechanisms, and monitoring are required because each technique addresses only a subset of the threat surface.

D. Insights and Limitations for FL

The majority of experimental work evaluates HFL on image-classification tasks under controlled non-IID partitions. Vertical FL threat models and defenses remain less mature. Balancing privacy, robustness to poisoning, communication efficiency, and accuracy under realistic system heterogeneity is an open systems challenge. Adaptive attackers that collude or target the defense mechanisms themselves are only beginning to be studied systematically.

IV. ADVERSARIAL ATTACKS AND DEFENSES IN TRANSFER LEARNING

A. Background in Transfer Learning

Transfer Learning reuses knowledge acquired on a source task or domain to accelerate learning on a

related target task. Common patterns include feature extraction from a pretrained backbone and fine-tuning of selected layers. Because high-quality source models are frequently obtained from public repositories or third-party providers, they introduce a supply-chain dimension absent from training from scratch.

B. Attacks on Transfer Learning

Backdoor and Trojan attacks planted in a source model can survive fine-tuning and transfer to the target task, especially when only a subset of layers is updated. The transferred backdoor may remain dormant on the target validation set yet activate when the trigger is presented at deployment. Adversarial examples crafted on a source model often transfer to the fine-tuned target—the well-known transferability phenomenon—enabling black-box attacks against downstream models. Model-extraction attacks recover approximate functionality of proprietary models through query access. Poisoning of source datasets can implant persistent vulnerabilities that are difficult for a downstream user to audit.

C. Defenses for Transfer Learning

Fine-pruning, Neural Cleanse, activation clustering, and related techniques attempt to detect or neutralize backdoors after a model has been transferred. Adversarial training and robust fine-tuning improve the resilience of the target model. Source-model verification, watermarking, and careful curation of pretrained weights reduce supply-chain risk. Certified robustness methods can be applied after transfer, although guarantees may weaken under distribution shift between source and target domains.

D. Insights for Transfer Learning

The convenience and performance benefits of large pretrained models must be weighed against the difficulty of fully auditing their training provenance. Transferability is a double-edged sword: it enables efficient black-box attacks but also allows a defense developed on one model to benefit another. Systematic evaluation across diverse source–target pairs remains relatively rare and is an important direction for future work.

V. ADVERSARIAL ATTACKS AND DEFENSES IN DEEP REINFORCEMENT LEARNING

A. Background in Deep Reinforcement Learning

Deep Reinforcement Learning combines deep function approximation with sequential decision-making. An agent interacts with an environment, receives observations, selects actions according to a policy, and obtains scalar rewards. Value-based methods (DQN), policy-gradient methods, and actor-critic algorithms are widely used. Because policies may control physical systems—vehicles, robots, drones—adversarial perturbations can have immediate safety consequences that go beyond classification error.

B. Attacks on Deep Reinforcement Learning

Observation (state) attacks add perturbations to sensor inputs so that the agent selects suboptimal or unsafe actions. Reward attacks manipulate the scalar feedback signal, causing the agent to learn an incorrect objective. Action-space attacks alter the action that is actually executed. Environment-dynamics attacks change transition probabilities or introduce non-stationarity. In multi-agent settings an adversary can train an adversarial policy that systematically exploits a victim agent without directly perturbing observations.

Attacks may target the training phase (poisoning), the deployment phase (attacking a fixed policy), or both. The temporal dimension means a successful attack need not succeed at every time step; a few critical perturbations can produce catastrophic trajectories.

C. Defenses for Deep Reinforcement Learning

Adversarial training of policies against perturbed observations improves empirical robustness at the cost of increased sample complexity. Robust reinforcement-learning formulations explicitly optimize performance under worst-case disturbances within a defined uncertainty set. Detection of anomalous observations or rewards, certified defenses applied to function approximators, and ensemble policies provide additional protection. Safety constraints, shielding layers, and runtime monitors can limit the damage residual attacks can inflict by overriding unsafe actions.

D. Insights for DRL

Reported robustness is highly sensitive to the choice of evaluation environment. Atari 2600 games and MuJoCo continuous-control tasks dominate the literature. Transfer of defenses developed for supervised DNNs to sequential DRL settings is non-trivial because of temporal credit assignment, non-stationary data distributions, and the need to preserve task performance under constraints. Safety-critical applications demand evaluation protocols that go beyond average reward to include worst-case performance and constraint-violation metrics.

VI. EVALUATION METRICS, DATASETS, AND TESTBEDS

A. Metrics

Attack success is commonly measured by attack success rate (ASR), absolute or relative drop in clean accuracy, and the magnitude of the perturbation (ℓp norms). For federated settings, additional metrics include the fraction of compromised clients required and the degradation of global-model accuracy. Privacy leakage is quantified by membership-inference advantage, inversion reconstruction error, or property-inference accuracy. Defense quality is reported as robust accuracy under adaptive attacks, residual privacy leakage, communication and computation overhead, and for DRL, constraint-violation frequency and recovery time.

B. Datasets and Application Domains

Image-classification experiments rely heavily on MNIST, CIFAR-10/100, and ImageNet. Federated learning studies employ FEMNIST, Shakespeare, and other federated benchmarks that better reflect realistic client heterogeneity. Deep reinforcement learning is dominated by the Atari 2600 suite and MuJoCo continuous-control tasks; autonomous-driving and robotics simulators appear in safety-oriented work. Text, graph, and multi-modal domains remain less comprehensively covered by systematic attack-and-defense evaluations.

C. Testbeds and Reproducibility

Reproducible, publicly available testbeds and standardized adaptive-attack evaluation protocols are still emerging. Many papers report only non-adaptive or limited-threat results, complicating fair comparison and potentially producing over-optimistic robustness claims. Community efforts toward shared leaderboards, open evaluation suites, and consistent reporting of threat models are of high value for the field.

Several open-source libraries facilitate reproducible adversarial research. Foolbox, CleverHans, and ART (Adversarial Robustness Toolbox) provide standardized implementations of common attacks and defenses across multiple deep-learning backends. For federated settings, PySyft and Flower support simulated FL environments. RLlib and Stable-Baselines3 are commonly used for DRL experiments. Standardizing threat specifications in benchmark suites remains an active area of community effort.

VII. COMMONALITIES AND DIFFERENCES ACROSS THE FOUR PARADIGMS

A. Commonalities

All four paradigms rely on deep, differentiable models and are therefore susceptible to gradient-based adversarial examples and to data or model poisoning. The transferability of adversarial examples appears across DNNs, transfer learning, and, to a lesser extent, deep reinforcement learning. Backdoor attacks surface in DNNs, federated learning, and transfer learning. Privacy leakage via gradients or model updates is a shared concern for federated learning and for collaborative or multi-agent reinforcement learning. In every case, adaptive adversaries that know the defense can substantially reduce reported robustness numbers.

B. Differences

Standard DNNs are typically centralized and process static inputs. Federated learning introduces distributed, potentially malicious participants and non-IID data distributions, making aggregation rules and client selection first-class security concerns. Transfer learning adds a supply-chain dimension through the reuse of pretrained source models; provenance and verification become critical. Deep reinforcement learning is sequential and interactive: attacks can target entire trajectories and the reward signal, not only a single static input, and defenses must preserve long-term task performance under constraints.

Consequently, defense strategies must be specialized. Robust aggregation and secure aggregation are central to FL. Source verification and careful fine-tuning are central to TL. Constrained or robust policy optimization together with runtime monitoring and shielding are central to DRL. Techniques that work well in one paradigm often require non-trivial adaptation before they become effective in another.

C. Cross-Cutting Lesson

No paradigm currently possesses a universal defense. Practical security requires defense-in-depth that combines robust training, anomaly detection, cryptographic or statistical privacy mechanisms, and continuous monitoring, all evaluated under realistic adaptive adversaries and system constraints. A defense that looks strong under a narrow threat model can fail when the attacker's knowledge, resources, or goals change.

TABLE I
Comparison of Attack and Defense Characteristics Across Paradigms

Feature	DNN	Fed. Learning	Transfer L.	DRL
Data	Centralized	Distributed	Source+Target	Interactive
Main Attack	Adv. Examples	Poisoning	Backdoor	Obs. Attack
Key Defense	Adv. Training	Robust Agg.	Fine-Pruning	Robust RL
Privacy Risk	Low	High	Medium	Medium
Certified	Yes (limited)	Emerging	Limited	Limited

VIII. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This survey has reviewed adversarial attacks and countermeasures for deep neural networks, federated learning, transfer learning, and deep reinforcement learning. We have summarized representative threat models, attack techniques, defense mechanisms, evaluation practices, and the relationships among the four paradigms. The analysis reveals a field that is rapidly maturing but still has significant gaps in standardization, scalability, and real-world evaluation. A recurring theme is the gap between adversarial robustness as measured in controlled laboratory settings and robustness as required in deployment. Published defense results are frequently evaluated under restricted threat models—non-adaptive adversaries, limited perturbation sets, or single-paradigm scenarios. When re-evaluated under stronger or more realistic attacks, many defenses provide substantially less protection than initially reported. Closing this credibility gap is arguably the most important short-term priority for the field.

A second recurring theme is the cost of robustness. Whether measured in computation (adversarial training), privacy-utility trade-off (differential privacy), communication overhead (secure aggregation), or sample complexity (robust RL), security improvements come at a price. Understanding and reducing these costs—especially for large-scale models and resource-constrained clients—is a persistent research challenge.

Looking forward, several research priorities stand out:

- Adaptive and standardized evaluation protocols that reflect realistic threat models, physical constraints, and system heterogeneity, so that robustness claims become comparable and trustworthy.
- Scalable robust-training methods that remain practical for large models and for non-IID federated settings, reducing the current gap between laboratory results and deployable systems.
- Principled combinations of privacy (differential privacy, secure aggregation) and robustness that make the resulting trade-offs explicit and quantifiable rather than implicit.
- Certified robustness guarantees that extend beyond small ℓp balls and static classification to sequential decision-making and multi-agent interactions.
- Supply-chain security for pretrained models, together with verification tools that are usable by non-expert downstream practitioners.
- Domain-specific benchmarks and testbeds for safety-critical applications such as autonomous driving, medical imaging, and industrial control.

• Deeper theoretical and empirical understanding of when and why attacks and defenses transfer across architectures, tasks, and paradigms.

Building trustworthy deep-learning systems will require sustained collaboration among machine-learning researchers, security experts, systems designers, and domain practitioners, guided by rigorous, reproducible evaluation. The four paradigms reviewed here—DNNs, FL, TL, and DRL—are increasingly deployed together in real systems, and their security properties interact in ways that single-paradigm analysis cannot capture. Cross-paradigm threat modeling and defense co-design therefore represent the frontier toward which the field should now direct significant attention.

REFERENCES

[1] H. Ali, D. Chen, M. Harrington, N. Salazar, M. Al Ameedi, A. F. Khan, A. R. Butt, and J.-H. Cho, "A Survey on Attacks and Their Countermeasures in Deep Learning: Applications in Deep Neural Networks, Federated, Transfer, and Deep Reinforcement Learning," *IEEE Access*, vol. 11, pp. 120095–120140, 2023. Crossref

[2] P. Kairouz et al., "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, nos. 1–2, pp. 1–210, 2021. Crossref

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

[5] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Security and Privacy*, 2017, pp. 39–57. Crossref

[6] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," arXiv:1708.06733, 2017. https://arxiv.org/abs/1708.06733

[7] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2020.

[8] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Yu, B. Liu, and P. S. Yu, "Privacy and robustness in federated learning: Attacks and defenses," *IEEE Trans. Neural Networks and Learning Systems*, 2022.

[9] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Proc. Network and Distributed System Security Symp. (NDSS)*, 2018.

[10] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," *ICLR Workshop on Trustworthy Machine Learning*, 2017.

[11] A. Gleave, M. Dennis, C. Wild, N. Kant, S. Levine, and S. Russell, "Adversarial policies: Attacking deep reinforcement learning," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2020.

[12] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 1310–1320.

[13] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE European Symp. Security and Privacy*, 2016. Crossref

[14] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018. Crossref

[15] N. Carlini et al., "Extracted training data from large language models," in *Proc. USENIX Security Symp.*, 2021.

[16] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–34, 2023.

[17] T. Chen, S. Liu, S. Chang, Y. Cheng, L. Amini, and Z. Wang, "Adversarial robustness: From self-supervised pre-training to fine-tuning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] A. S. Rakin, Z. He, and D. Fan, "Bit-flip attack: Crushing neural network with progressive bit search," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019. Crossref

[19] A. Szegedy et al., "Intriguing properties of neural networks," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2014.

[20] Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805–2824, 2019. Crossref

[21] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22] A. Shafahi et al., "Poison frogs! Targeted clean-label poisoning attacks on neural networks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[23] A. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep text classification can be fooled by long and disguised adversarial examples," in *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2018.

[24] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS)*, 2017. Crossref

[25] M. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[26] Y. Sun, A. Bhowmick, and J. M. Karbasi, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2017.

[27] T. Lillicrap et al., "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2016.

[28] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015. Crossref

[29] T. Everitt, G. Lea, and M. Hutter, "AGI safety literature review," arXiv:1805.01109, 2018. https://arxiv.org/abs/1805.01109

[30] N. Papernot et al., "Semi-supervised knowledge transfer for deep learning from private training data," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.