

The Evolution of Adversarial Machine Learning Research

1 Introduction:

Machine learning systems have achieved remarkable success in various domains, including autonomous systems, healthcare, natural language processing, cybersecurity, and computer vision. However, this success has come with different vulnerabilities that challenge their reliability and security. Adversarial Machine Learning is an emerging field studying the attacks against machine learning models and defences to protect them. This is a critical field emerging at the intersection of machine learning and security.

The research has risen dramatically over the last decade, with numerous papers demonstrating powerful algorithmic attacks followed by papers discovering advanced defences. Yet, a troubling disconnect has emerged; while academic research has discovered sophisticated attacks mostly based on mathematical computations of gradients, real-world evidence suggests that attackers employ fairly simple tactics to subvert ML-driven driven due to the computational, knowledge and resource limitations [1, 15].

We trace the evolution of adversarial machine learning research and establish a foundation understanding through increasingly complex attacks and defences, highlighting theoretical advancements, gradient-based methods, and shaping the research landscape. Understanding the trajectory helps us contextualise subsequent analysis of recent publications. Then, we evaluate the extent to which research has evolved and if it has addressed the practical security considerations relevant to industry deployments, revealing a critical gap between theoretical advancements and industry deployments.

2 The Emergence of Adversarial Examples:

The modern era of adversarial machine learning began with the discovery of intriguing properties of neural networks by Szegedy et al (2013) [33]. This revealed that imperceptible tiny perturbations in the input data could fool state-of-the-art deep learning models to produce incorrect outputs with high confidence. Importantly, these perturbations weren't random errors of ML systems, but rather a deep vulnerability that could transfer across different models, undermining the security of any machine learning system. This transferability means that adversarial examples crafted for one model often fool other models with different architectures, even when those models were trained independently, a property that makes these attacks particularly dangerous. This is a fundamental weakness of neural network architecture and drew a lot of attention and subsequent research.

Goodfellow et al (2014, 2015) [10, 11] proposed that the linear nature of neural networks, rather than the non-linearity or overfitting, was the primary cause of adversarial vulnerability. They introduced a Fast Gradient Sign Method (FGSM), simplifying the process of adversarial input generation with just a single step. The simplicity and effectiveness of FGSM established a pattern that eventually dominated the subsequent research. Gradient-based attacks became the standard.

These foundational works shaped the trajectory of adversarial ML research. They assumed that adversaries had a white box access to the models, including architecture, details and gradients. Second, they focused on imperceptible perturbations. Third, they emphasised mathematical optimisation as a natural way of computing the gradient and the attack vector. While these frameworks enabled rigorous analysis, they shaped the research away from the messier realities of real-world attack scenarios.

3 The Arms Race: Increasingly Sophisticated Gradient Attacks

Following the initial ideas, researchers started to discover increasingly sophisticated gradient-based attacks, each one increasing in computational complexity and mathematical algorithms.

Moosavi-Dezfooli et al (2015, 2016) [23, 24] introduced DeepFool, which computed the minimum perturbations iteratively, making the classification cross decision boundaries. This is a great improvement to the FGSM as the DeepFool requires multiple computations, making the adversarial perturbations even smaller and precise. This approach further embeds gradient computation as a foundation for adversarial ML research.

Carlini and Wagner (2017) [6] discovered a C and W attack formulating adversarial examples as a complex optimisation problem tailored to distance metric (L0, L2 and Linf). These distance metrics measure different types of perturbation: L0 counts the number of changed pixels, L2 measures Euclidean distance, and Linf measures the maximum per-pixel change. This involved solving the optimisation problem where the formulation minimizes the perturbation distance while ensuring misclassification, balancing attack success against perturbation visibility through careful hyperparameter selection.

This approach also required a significant gradient computation, hyperparameter tuning and computational resources. It achieved numerous successes. At that time, defensive distillation (Papernot et al, 2015, 2016) [28, 29] was the standard defence. Carlini and Wagner demonstrated that their attacks succeed against many proposed defence methods, including the defensive distillation.

This led to an arms race between attacks and defence research. Defences were proposed and evaluated against known attacks, but were subsequently broken by sophisticated gradient-based adaptive attacks. This was a never-ending process of mathematical sophistication, with each iteration of research requiring a deeper gradient analysis and more computational resources. While theoretical frameworks became more robust and concrete, the adversarial ML research diverged from practical attack scenarios where adversaries face computational constraints, limited model access, and different cost-benefit calculations than assumed in academic threat models.

4 Physical World Attacks:

Researchers extended gradient-based methods to physical world scenarios. Kurakin et al (2016) [16] demonstrated that adversarial examples printed and photographed could fool classifiers, but their approach still relied on gradient-based optimisation, modified to simulate camera capture and environment variables.

Jia et al (2022) [14] advanced physical adversarial examples against autonomous vehicle traffic sign recognition systems. Building on earlier work in physical attacks, they tackled the challenge of creating adversarial perturbations that remain effective under real-world conditions like varying lighting, angles, and distances. Their method was:

- Extend image transformation distributions with blur and resolution transformations
- Bounding box filters for perturbation efficiency
- gradient-based method for optimising perturbations

They demonstrated the real-world relevance by successfully fulfilling a 2021 vehicle's traffic sign recognition system. While the attack was successful, a natural question is would actual attackers targeting autonomous vehicles employ such computationally intensive gradient-based methods, or would simpler approaches like physically obscuring signs or using adversarial stickers suffice?

Chahe et al (2023) [7] introduced dynamic adversarial attacks using screens on moving vehicles employing a Screen Image Transformation Network trained through gradient-based optimisation. The approach assumed capabilities like extensive computational resources, gradient access, ability to train a specialised network that may not align with practical attackers' resources or motivations.

5 Black Box Attacks

Realising that white-box attacks were unrealistic in many scenarios due to a lot of assumptions about the internal workings of the models, researchers developed black-box strategies. Papernot et al (2016, 2017) [26, 27] introduced substitute model training. The adversary queries the target model and receives output, typically class predictions or probability scores. Based on these outputs, the adversary trains a local substitute model on these labelled samples, and generates adversarial examples using gradients from the substitute network architecture and attacks the original system using these examples.

This approach showed a remarkable success, which misclassified 84% for MetaMind's API, 96% against Amazon and 89% against Google. This demonstrates a high practical implication. However, the method still relied on the gradient computation. While more realistic than white box assumptions, this approach still assumed adversaries possessed significant technical sophistication and abundant resources. The question of whether real attackers without ML expertise might employ non-gradient-based approaches is still not answered well.

6 Defence Mechanisms: The Gradient Masking Trap

Numerous studies have shown powerful defences against adversarial attacks. They largely mirror gradient-based methods and advanced mathematics.

6.1 Detection-Based Defences

Early research tried to detect adversarial examples through statistical properties or learned features. Metzen et al (2017) [21] suggested detector subnetworks. Grosse et al (2017) [12] explored statistical tests based on distributional differences. Meng and Cheng (2017) [20] introduced MagNet, which combined detector and reformer networks.

Carlini and Wagner (2017) [5] systematically failed existing 10 state-of-the-art defences, demonstrating that all could be bypassed through adaptive attacks with a modified loss function. This revealed a fundamental problem: defence evaluated only for known attacks failed when adversaries adapted their gradient-based optimisation to account for the defence mechanism. This established that detection-based defence was significantly harder to implement effectively than previously anticipated.

6.2 Adversarial Training with Gradient-Based Framework:

Adversarial training emerged as an important defence direction. Goodfellow et al (2014, 2015) [10, 11] introduced the concept: augment training data with adversarially perturbed inputs generated via gradient-based methods. Madry et al (2017, 2018) [19, 25] reformulated the adversarial robustness problem through robust optimisation through a min-max framework. The min-max formulation seeks to minimize loss over the worst-case adversarial perturbations within a bounded set, essentially training the model to perform well even against the strongest attacks it might face within specified constraints.

The formulation required finding the worst-case adversarial perturbation via gradient ascent for each sample during each training epoch. Tramer et al (2017) [35] introduced Ensemble Adversarial Training, which improved black-box robustness by training on perturbations transferred from multiple models.

The pattern of research is clear: defences were designed to counter gradient-based attacks, evaluated against gradient-based attacks, and optimised assuming adversaries would use gradient-based methods. This created a self-reinforcing cycle where both attacks and defences revolved around the same gradient-based paradigm, potentially neglecting simpler attack vectors that real-world adversaries might implement.

6.3 Certified Robustness

Research has actively looked for formal guarantees that models remain correct within specific perturbation bounds. Li et al (2020) [17] systematically certified robustness. However, these methods came with severe accuracy-robustness tradeoffs and required significant computational resources.

Weber et al (2020) [37] extended certification to blackdoor attacks through RAB, using randomised smoothing to provide provable guarantees. However, certified guarantees offer lower robust accuracy than empirical adversarial training. This makes it deployment challenging in resource-constrained environments.

Certified robustness was highly rigorous and theoretically complex. The question of whether such sophisticated guarantees were necessary or cost-effective for real-world deployment mostly remained unaddressed.

7 Privacy Attacks

While evasion attacks were prominent in early research, the field expanded to capture that ML models could leak sensitive information about training data, representing distinct privacy violations.

7.1 Membership Inference

Membership inference attacks test whether a given data point was used in model training. Ye et al (2021) [38] developed enhanced attacks based on comprehensive hypothesis testing, building upon earlier membership inference work by introducing more sophisticated statistical methods to detect training set membership with higher accuracy. Carlini et al (2022) [3] argued for evaluating attacks at very low false positive rates. Their likelihood ratio attack achieved 10x higher power by carefully combining statistical signals from reference models. The attacks assumed adversaries could query models extensively and possessed statistical expertise to interpret results, potentially overestimating attacker sophistication compared to real-world scenarios.

7.2 Property Inference

Property inference attacks extract global training data statistics. Chase et al (2021) [8] studied poisoning-enhanced property inference, demonstrating that adversaries controlling portions of training data could boost inference success. Chaudhary et al (2022) [9] introduced SNAP for efficient property extraction through poisoning.

Balle et al (2022) [2] study training data reconstruction by informed adversaries, deriving closed-form attacks for convex models and training reconstructor networks. Carlini et al (2023) [4] extracted over a thousand training examples from diffusion models, demonstrating that these models were substantially less private than previous generative models.

7.3 Federated Learning Privacy

Privacy concerns intensified in distributed learning. Pasquini et al (2021) [30] showed that secure aggregation, designed specifically to protect privacy in federated learning, could be exploited by taking advantage of model inconsistencies. Pasquini et al (2022) [31] extended this to peer-to-peer decentralised learning, demonstrating that decentralisation also doesn't guarantee security advantages. Salem et al (2022) [32] provided unified game-based frameworks for systematising privacy attacks, bringing theoretical rigour to privacy analysis.

8 Training Time Attacks: Poisoning and Backdoors

Training time attacks manipulate the learning process itself, representing threats distinct from test-time evasion attacks.

8.1 Backdoor Attacks in Self-Supervised Learning

Self-supervised learning, where models learn representations from unlabeled data before being fine-tuned for specific tasks, has become increasingly popular. However, this paradigm introduces new attack surfaces. Jia et al (2021) [13] introduced BadEncoder, demonstrating backdoor injection into pre-trained encoders that propagated to downstream tasks. Their approach included:

- access to pre-training data or the ability to influence the pre-training process
- careful construction of poisoned samples and triggers
- optimisation to ensure backdoor persistence across fine-tuning

Tao et al (2024) [34] advanced this approach with Drupe, a distribution-preserving backdoor attack that transformed poisoned samples to appear in distribution, evading detection based defences. Liu et al (2024) [18] showed through PreCurios that pre-trained models could serve as privacy traps as they can propagate privacy risks when fine-tuned to downstream tasks based on the badly encoded model.

8.2 Advanced Poisoning Strategies

Tramer et al (2022) [36] introduced "truth serum" attacks where just poisoning less than 0.1% of training data boosted the inference attack performance by 1x or 2x. However, the practical implication of this is still a question. In what real-world scenario do adversaries control such a precise fraction of training data? What are the economic incentives of such attacks?

9 The Growing Disconnect with the Industry

As the theoretical sophistication of adversarial ML research grew, concerns emerged about whether this work addressed real-world security needs. Several studies began examining this research-practice gap directly.

Kumar et al. (2020) [15] conducted interviews with 28 organisations, finding that practitioners lacked tactical and strategic tools for adversarial ML threats. Grosse et al. (2024) systematically quantified this disconnect by surveying 271 industrial practitioners, revealing that while academic threat models were theoretically applicable, research consistently overestimated attacker capabilities particularly regarding training data access and query budgets. Critically, they found that simpler attacks requiring fewer resources (e.g., ≤ 100 queries) could target 28.8-44.4% of deployed models, yet received far less research attention than complex gradient-based methods. Mink et al. (2023) [22] provided complementary insights through interviews with 21 data scientists and engineers, identifying three primary barriers to defence deployment: lack of institutional motivation and educational resources, inability to adequately assess AML risk, and organisational structures prioritising other objectives over security. Practitioners often viewed security as outside their expertise: "Security is not my field, I'm a stats guy" revealing fundamental misalignments between ML role definitions and security responsibilities.

Apruzzese et al crystallised this gap in the 2023 paper "Real Attackers Don't Compute Gradients" [1]. Through real-world case studies and a comprehensive examination of adversarial ML

papers from the top 4 security conferences, they documented a fundamental misalignment. Attackers use simple tactics to fool ML systems, while research focuses on sophisticated gradient attacks. Their case study reveals an important distinction about real-world ML system compromises:

- Simple input manipulation, like putting a false logo (without gradients)
- Exploiting system vulnerabilities rather than model-specific weaknesses
- Taking advantage of operational gaps more than algorithmic weaknesses
- Leveraging social engineering

This historical review reveals that adversarial machine learning is a field of substantial theoretical advancement but questionable practical implications. From foundational discoveries [10,11,33] through gradient attacks [6,23,24] and defences [17,19,25]. The field has grown to include privacy attacks [3,38], training time threats [13,34,36], and physical world scenarios [7,14]. The major research has focused on gradient-based methodologies, algorithmic vulnerabilities, threat models, and evaluation paradigms, prioritising worst-case theoretical scenarios over realistic attack economics.

Meanwhile, evidence from industry suggests that actual ML systems compromise follow different patterns, with attackers exploiting operational gaps and economic incentives over algorithmic sophistication.

This disconnect motivates the subsequent systematic analysis of recent adversarial ML publications. Building on this historical foundation, we examine papers published in top-tier conferences from 2022 to 2025 to evaluate where the current research is addressing the research-practice gap or is continuing the same historical patterns. Our analysis assesses the extent to which current work incorporates realistic threat models, providing insights into the direction it's heading towards. Through this examination, we aim to quantify the gap between adversarial ML research priorities and industry security needs, proposing directions for increasing the real-world impact of future research in this domain.

References

- [1] G. Apruzzese, H. S. Anderson, S. Dambra, D. E. Freeman, F. Pierazzi, and K. Roundy. "real attackers don't compute gradients": Bridging the gap between adversarial ml research and practice. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023.
- [2] B. Balle, G. Cherubin, and J. Hayes. Reconstructing training data with informed adversaries. In *IEEE Symposium on Security and Privacy*, 2022.
- [3] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy*, 2022.
- [4] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Schwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. *ArXiv.Org*, 2023.
- [5] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, 2017.
- [6] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017.
- [7] A. Chahe, C. Wang, A. Jeyapratap, K. Xu, and L. Zhou. Dynamic adversarial attacks on autonomous driving systems. *ArXiv.Org*, 2023.
- [8] M. Chase, E. Ghosh, and S. Mahlouljifar. Property inference from poisoning. In *IEEE Symposium on Security and Privacy*, 2021.
- [9] H. Chaudhari, J. Abascal, A. Oprea, M. Jagielski, F. Tramèr, and J. Ullman. Snap: Efficient extraction of private properties with poisoning. *ArXiv.Org*, 2022.
- [10] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2014.
- [11] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.
- [12] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. On the (statistical) detection of adversarial examples. *ArXiv: Cryptography and Security*, 2017.
- [13] J. Jia, Y. Liu, N. Z. Gong, Gong Yupei, and N. Zhenqiang. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. *ArXiv (Cornell University)*, 2021.
- [14] W. Jia, Z. Lu, H. Zhang, Z. Liu, J. Wang, and G. Qu. Fooling the eyes of autonomous vehicles: Robust physical adversarial examples against traffic sign recognition systems. In *Proceedings 2022 Network and Distributed System Security Symposium*, 2022.
- [15] R. S. S. Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, and S. Xia. Adversarial machine learning-industry perspectives. In *2020 IEEE Security and Privacy Workshops (SPW)*, 2020.

- [16] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2016.
- [17] L. Li, L. Li, T. Xie, and B. Li. Sok: Certified robustness for deep neural networks. In *IEEE Symposium on Security and Privacy*, 2020.
- [18] R. Liu, T. Wang, Y. Cao, and L. Xiong. Precurious: How innocent pre-trained language models turn into privacy traps. *ArXiv.Org*, 2024.
- [19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2017.
- [20] D. Meng and H. Chen. Magnet: A two-pronged defense against adversarial examples. In *Computer and Communications Security*, 2017.
- [21] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. *ArXiv: Machine Learning*, 2017.
- [22] Jaron Mink et al. "security is not my field, i'm a stats guy": A qualitative root cause analysis of barriers to adversarial machine learning defenses in industry. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2023.
- [23] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *Computer Vision and Pattern Recognition*, 2015.
- [24] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. 2016.
- [25] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv (Cornell University)*, 2018.
- [26] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *ACM Asia Conference on Computer and Communications Security*, 2016.
- [27] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *Computer and Communications Security*, 2017.
- [28] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, 2015.
- [29] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. 2016.
- [30] D. Pasquini, D. Francati, and G. Ateniese. Eluding secure aggregation in federated learning via model inconsistency. In *Conference on Computer and Communications Security*, 2021.

- [31] D. Pasquini, M. Raynal, and C. Troncoso. On the (in)security of peer-to-peer decentralized machine learning. In *IEEE Symposium on Security and Privacy*, 2022.
- [32] A. Salem, G. Cherubin, D. Evans, B. Köpf, A. Paverd, A. Suri, S. Tople, and S. Zanella-Béguelin. Sok: Let the privacy games begin! a unified treatment of data inference privacy in machine learning. In *IEEE Symposium on Security and Privacy*, 2022.
- [33] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2013.
- [34] G. Tao, Z. Wang, S. Feng, G. Shen, S. Ma, and X. Zhang. Distribution preserving backdoor attack in self-supervised learning. In *IEEE Symposium on Security and Privacy*, 2024.
- [35] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2017.
- [36] F. Tramèr, R. Shokri, A. S. Joaquin, H. M. Le, M. Jagielski, S. Hong, and N. Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. In *Conference on Computer and Communications Security*, 2022.
- [37] M. Weber, X. Xu, B. Karlaš, C. Zhang, and B. Li. Rab: Provable robustness against backdoor attacks. In *IEEE Symposium on Security and Privacy*, 2020.
- [38] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri. Enhanced membership inference attacks against machine learning models. In *Conference on Computer and Communications Security*, 2021.