IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14614

Adversarial Attacks on AI System, Vulnerabilities, Threat Models and Defensive Implications

Abdullateef Ajibola Adepoju[1], Saidu Sunbo Akanji[2], Rukayya Abdulganiyu Adepoju[3], Atuma Ochenu Lawrence[4]

Randatech Innovative Insititute.No 45 Zaria road, Dangi underpass,Tarauni, Kano, Kano State[1]

Department of Electrical & Electronics Technical Education, Federal College of Education (Technical)Yauri[2]

Department of Pure and Industrial Chemistry, Bayero University Kano, Nigeria[3]

Department of Science Education, Federal University Dutsinma, Katsina State[4]

Corresponding Email: saiduakanji@gmail.com

Abstract: Concerns about the security of artificial intelligence systems have grown sharply as these technologies take on increasingly consequential roles in healthcare, infrastructure, finance and national security. One of the most pressing threats in this space is the adversarial attack, an intentional, engineered input designed to cause an AI model to behave in ways its designers never intended. This paper brings together a decade of published evidence through a structured meta-analysis of secondary adversarial attack case studies spanning the period 2015 to 2025. The domains covered include computer vision, natural language processing, cybersecurity tools, autonomous systems and decision-support platforms. What emerges from this synthesis is not a collection of isolated incidents but a consistent picture, adversarial weaknesses are baked into how modern machine learning systems are built, stemming from their sensitivity to high-dimensional inputs, poorly defined threat assumptions, and exposure at multiple points along the data supply chain. Among the attack types reviewed, evasion attacks appeared most frequently, accounting for 78 percent of documented cases, while backdoor and data poisoning attacks, though rarer, often left the most lasting damage. One of the more striking findings is how readily attack strategies move across domains and model types and how closely AI security threats are beginning to resemble traditional cybersecurity problems. Defences, meanwhile, have struggled to keep pace, most of the mitigation strategies reviewed broke down once attackers adapted. The paper concludes with a call for threat modelling that spans the full AI development lifecycle, evaluation methods that measure genuine resilience rather than clean-data accuracy and governance structures that treat adversarial robustness as a first-class requirement.

Index Terms: Adversarial Attacks, Machine Learning Security, Evasion Attacks, Data Poisoning, AI Robustness, Threat Modelling, Defensive AI.

1. INTRODUCTION

1.1 Background and Context

Not long ago, the question of whether an AI system could be deliberately tricked into making dangerous errors was largely a theoretical one. That is no longer the case. Machine learning models today sit at the centre of decisions that affect people's lives, such as screening job applicants, diagnosing disease, detecting fraud, guiding autonomous vehicles and filtering content at a scale no human team could match. As these systems have become embedded in critical infrastructure, the question of what happens when a determined adversary targets them has become impossible to ignore [3,4].The core problem is this: AI models, particularly deep neural networks, can be fooled by inputs that look perfectly normal to a human but cause the model to behave erratically or dangerously. These are adversarial attacks, and they have been demonstrated in enough real and near-real settings to move beyond theoretical concern. Traffic sign classifiers have been fooled by subtle sticker placements. Malware has slipped past classifiers undetected. Content moderation tools have been manipulated into allowing harmful material through [5-7]. Each of these examples points to a gap between how AI systems perform in controlled tests and how they hold up when someone is actively trying to break them.

1.2 Problem Statement and Scope

The uncomfortable truth is that despite years of dedicated research, adversarial attacks on AI remain an open and largely unsolved problem. Proposed defences, from adversarial training to input sanitisation, have shown real but limited promise

DOI: 10.17148/IJIREEICE.2026.14614

, they tend to work well against the attacks they were designed to counter and fail when attackers adapt [8,9]. Much of this defensive work has also been carried out in lab settings that do not reflect the messiness of real deployment, making it hard to know how much protection these techniques actually offer in practice. Compounding this, the research landscape is fragmented. Studies tend to focus on a single domain or attack type, which means that patterns visible across the wider literature , recurring weaknesses, shared attacker strategies, structural flaws in how AI systems are built and maintained, go largely unnoticed [10]. This paper takes a different approach. By drawing together secondary case studies from across domains and threat models, it aims to surface the commonalities that individual studies cannot see. Adversarial attacks are categorised here by attacker knowledge (white-box versus black-box), by when they occur (training time versus inference time) and by their delivery mechanism (digital versus physical). Across all these categories, one finding stands out, attack strategies are becoming increasingly transferable across models and domains [12,13].

1.3 Study Aim and Contribution

This research does not set out to introduce new attacks or build new defences. Its purpose is more foundational, to take stock of what the accumulated evidence actually tells us about how and why AI systems fail under adversarial pressure. By synthesising secondary case studies systematically rather than reviewing them selectively, the analysis identifies structural failure modes that transcend any particular model architecture or application context. These findings carry practical implications for how AI systems are designed, how their robustness is evaluated and how regulators and developers should think about adversarial risk as a governance responsibility rather than a purely technical one.

2. MATERIALS AND METHODS

2.1 Study Design and Case Selection

This study uses a meta-analytical design, bringing together qualitative comparison and quantitative trend analysis across a corpus of documented adversarial attack case studies. Each individual case study, whether a controlled experimental demonstration or a documented near-real-world incident, serves as the unit of analysis. Literature was gathered from IEEE Xplore, the ACM Digital Library, SpringerLink, ScienceDirect and arXiv, using search terms that included adversarial attack, adversarial example, machine learning security, model poisoning, evasion attack and AI robustness. Reference lists from major surveys and influential individual papers were also reviewed manually to reduce the risk of missing relevant work. A case study was included if it clearly described an adversarial attack on a supervised or semi-supervised AI system, provided enough methodological detail to understand what the attack involved and under what conditions it was tested, and was published between 2015 and 2025 in either a peer-reviewed venue or a preprint with substantial citation impact. Studies that remained purely theoretical or that addressed only defensive proposals without demonstrating an actual attack, were excluded. Where a single case study illustrated more than one attack type, it was counted once per category in the frequency analysis.

2.2 Data Extraction and Analytical Framework

For each selected case study, information was extracted on the AI application domain, model architecture, learning paradigm, the level of knowledge the attacker was assumed to have, the type of attack, the evaluation setting and the impact reported on system behaviour. Where available, contextual information, including deployment assumptions, the attacker's likely motivation and any defensive measures already in place was also recorded. This dual focus on technical and contextual detail was deliberate, attack viability in the real world often depends less on model-level accuracy figures and more on the conditions surrounding a deployment [9,10]. Coding was carried out iteratively, with category definitions refined as the corpus was reviewed and anchored throughout to establish adversarial taxonomies to keep subjectivity in check [12,16]. Rather than attempting to aggregate performance metrics across studies , which use incompatible tasks, datasets and measurement scales, the analysis focuses on attack outcomes that can be compared meaningfully across contexts, such as mis-classification rates, successful system evasion and disruption of control behaviour. Publication bias and temporal bias were both acknowledged as limitations and managed by emphasizing recurring cross-study patterns over specific numerical claims [8,14].

3. RESULTS AND ANALYSIS

3.1 Distribution of Case Studies Across Domains and Time

Computer vision is by some distance the most heavily studied domain in the adversarial attack literature, with 26 case studies in the corpus. Attacks on image classifiers and object detection systems, particularly those used in traffic sign recognition and biometric identification , account for the bulk of this work [3,5,14]. Natural language processing is the next most targeted area, where documented attacks on sentiment analysis tools, content moderation systems and question-answering models have made clear that text-based AI is no more inherently secure than its vision-based counterpart [7].

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14614

Cybersecurity applications, especially malware and intrusion detection, form a natural third cluster in which the adversarial contest between attacker and defender is not an abstraction but an ongoing operational reality [6,10]. Autonomous systems and decision-support platforms appear less frequently in the literature but carry out-sized practical stakes.

Table 1: Distribution of adversarial attack case studies across AI application domains

AI Application Domain	No. of Case Studies	Dominant Attack Types	Typical Threat Model
Computer Vision	26	Evasion, Physical attacks	White-box, Black-box
Natural Language Processing	14	Evasion, Trigger-based	Black-box
Cybersecurity (Malware/IDS)	11	Evasion, Poisoning	Adaptive attacker
Autonomous / RL Systems	6	Reward manipulation, Evasion	Partial observability
Decision Support Systems	3	Poisoning, Inference	Insider / Data-level

The publication timeline tells its own story. Adversarial attack research grew steadily through 2015 and 2016 but accelerated sharply after 2017, which is roughly when deep learning became the dominant paradigm across industry and academia and when purpose-built attack toolkits became widely available. Early papers were largely proof-of-concept demonstrations [3,4]; the more recent literature increasingly grapples with questions of scalability, real-world transferability and what adversarial risk actually looks like once a system has been deployed [13,19].

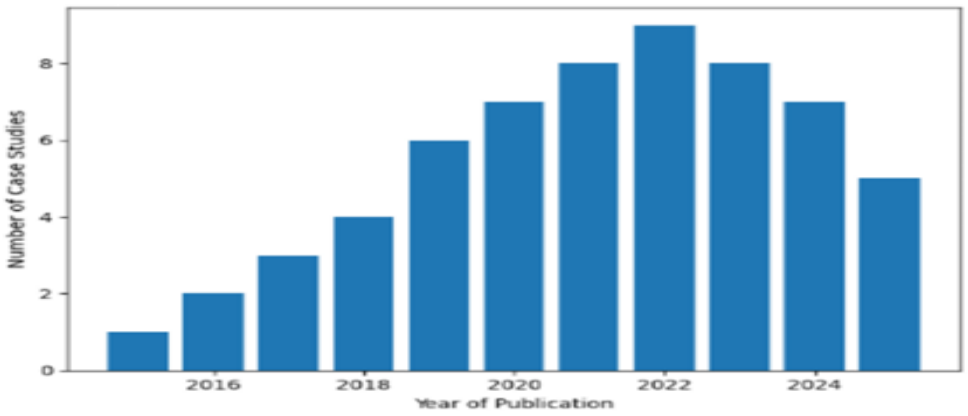

Fig 1. Temporal distribution of adversarial attack case studies (2015-2025)

3.2 Prevalence of Adversarial Attack Types

Evasion attacks dominate the corpus , they appear in 78 percent of case studies and it is not difficult to see why. These attacks operate at inference time, manipulating the inputs a model receives rather than the model itself, and they work across a wide range of applications [4,6,7]. Crucially, they can often be carried out with limited knowledge of the target system, making them accessible to a broad range of adversaries. Backdoor and data poisoning attacks tell a different story. They are less common in the literature, appearing in 32 and 25 percent of cases respectively, but the damage they cause tends to be harder to detect and address. A model that has been poisoned during training may pass every standard quality check, its performance on clean test data looks fine , while quietly failing in exactly the ways an attacker intended whenever a specific trigger is present [10,15]. Model extraction and membership inference attacks round out the picture, targeting not the model's predictions but the privacy of its training data or the confidentiality of its architecture [16].

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14614

Table 2: Frequency of adversarial attack categories across analysed case studies

Attack Category	Frequency (%)	Lifecycle Stage	Representative Domains
Evasion	78%	Inference-time	Vision, NLP, Malware
Data Poisoning	32%	Training-time	Malware, Decision Systems
Backdoor	25%	Training-time	Vision, NLP
Model Extraction	18%	Post-deployment	API-based models
Inference / Membership	15%	Post-deployment	Healthcare, NLP

Figure 2: Taxonomy of adversarial attack types and threat models

3.3 Patterns of Systemic Vulnerability

One of the most important things this meta-analysis reveals is that adversarial vulnerabilities are not random or model-specific, they keep appearing in the same forms, across different architectures and application areas, because they reflect fundamental properties of how machine learning systems are built. The most persistent of these is what might be called high-dimensional fragility. Modern classifiers learn to navigate extremely complex input spaces and in doing so they create decision boundaries that are tight and sensitive. A small, targeted perturbation that would not register as meaningful to any human observer can push an input across one of these boundaries and produce a completely different prediction [3,14]. This is not a bug in a specific model; it is a consequence of the way statistical learning works, and it shows up in convolutional networks, recurrent models and transformers alike. A second recurring pattern is the failure to think carefully about who might be trying to attack a system when it is being designed. Many of the case studies in this corpus describe successful attacks that exploited assumptions the development team had simply never questioned, assumptions about who would interact with the model, what inputs it would receive and what an adversary would be motivated to do [9,18]. Finally, a cluster of vulnerabilities tied specifically to the deployment phase emerged from the data: models updated without monitoring, training data ingested from unverified sources, third-party components integrated without security review [10,15]. These findings make clear that adversarial risk does not live only inside the model, it is distributed across the entire lifecycle of an AI system.

Table 3: Cross-domain vulnerability patterns observed in adversarial attack case studies

Vulnerability Pattern	Description	Domains Observed
High-dimensional sensitivity	Small perturbations causinglarge disproportionate output shifts	Vision, NLP
Weak threat modelling	Unrealistic attacker assumptions which undermine defensive design	All domains
Data pipeline exposure	Untrusted or poisoned training data embeds latent failures	Cybersecurity
Deployment oversights	Absent monitoring, auditing or update governance post-deployment	Autonomous systems

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026
DOI: 10.17148/IJIREEICE.2026.14614

3.4 Severity of Attack Outcomes and Defensive Measures

The consequences of successful adversarial attacks range from modest and recoverable to severe and systemic. At the serious end, adversarial perturbations applied to traffic sign classifiers have caused persistent miss-identification of objects that autonomous vehicles rely on to navigate safely [5]. In cybersecurity applications, evasion attacks have rendered malware detection systems functionally useless against specially crafted payloads [6]. Backdoor attacks sit in a category of their own when it comes to danger,because the compromised model continues to perform normally on clean inputs, these failures can go undetected for extended periods, surfacing only when an attacker deliberately activates the trigger condition [15]. Several case studies also document something worth dwelling on , defences that appeared effective in testing were bypassed once an attacker had the opportunity to adapt [9,19]. This suggests that passing a robustness evaluation is not the same as being robust. Adversarial training remains the most widely adopted mitigation approach, and it does provide meaningful protection in some circumstances , but it is expensive to implement, degrades performance on clean data, and tends to be effective only against the specific attack types it was trained to handle [8,14]. Other strategies, including input filtering and gradient obfuscation, have shown mixed results at best and have been broken outright in a number of documented cases [9]. Perhaps the most telling gap in the defensive literature is the near-complete absence of evaluations conducted under realistic operating conditions with limited computational budgets, evolving attackers and defenders who do not have perfect information about what they are defending against.

Table 4: Defensive approaches reported in adversarial attack case studies

Defence Strategy	Reported Effectiveness	Common Limitations
Adversarial Training	Moderate	High cost; attack-specific
Input Preprocessing	Low to Moderate	Readily circumvented adaptively
Detection Mechanisms	Inconsistent	Elevated false-positive rates
Gradient Obfuscation	Low	Fails under adaptive attack

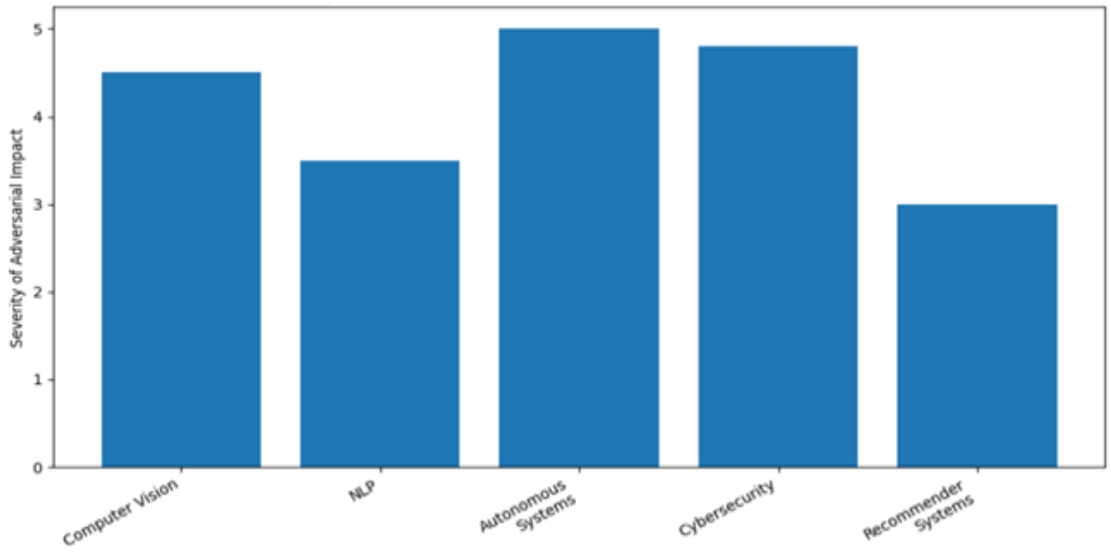

Figure 3: Severity of adversarial impact across AI domains

4. DISCUSSION

4.1 Interpretation and Cross-Domain Insights

The concentration of adversarial case studies in computer vision, autonomous systems and security applications is sometimes attributed to researcher preference or data availability. But looking at the evidence more carefully, it reflects something more fundamental, these are domains where the architectural choices that make deep learning so powerful also create the sharpest exposure to adversarial manipulation. Vision systems, for instance, operate in continuous, high-

DOI: 10.17148/IJIREEICE.2026.14614

dimensional input spaces and rely on learned representations whose decision surfaces are inherently sensitive to small, structured changes [23,24]. This is not incidental; it follows from the way these models are trained. What is perhaps more surprising is how consistently attack strategies transfer across domains that seem, on the surface, quite different from one another. The gradient-based methods first used to attack image classifiers have been adapted for language models, applied to control systems and repurposed against malware detectors with relatively little modification [27,28,29]. This transferability is not a coincidence. It reflects the fact that adversarial attacks exploit statistical properties that are common to machine learning systems generally, not quirks of any particular architecture or task. The implication is significant, a team securing an NLP system should be paying close attention to what is being learned in computer vision adversarial research, and vice versa.

Another pattern worth noting is the growing resemblance between AI security threats and the kind of attacks that cybersecurity professionals have been dealing with for decades. Model extraction is reverse engineering by another name. Membership inference is a form of side-channel attack. API abuse in the context of AI systems mirrors techniques long used against conventional software [30]. This convergence argues strongly for integrating AI security into existing security frameworks rather than treating it as a separate discipline with its own bespoke toolbox. It also argues for more dialogue between the AI research community and the cybersecurity community, two groups that, at present, largely operate in parallel. One more finding deserves emphasis: the persistent gap between how robustness is measured in research and what robustness actually means in a deployed system. Models that perform well on standard adversarial benchmarks have repeatedly been shown to be brittle when exposed to adaptive attackers or conditions that differ even modestly from those benchmarks [26]. This is not a minor calibration issue, it points to a systemic problem in how the field defines and tests security.

4.2 Engineering, Policy and Governance Implications

For engineers building AI systems, the clearest takeaway from this analysis is that threat modelling cannot be an afterthought. A substantial portion of the failures documented in the reviewed case studies were not the result of sophisticated new attack techniques , they exploited gaps that a realistic threat model would have identified before deployment [31]. The implication is practical, teams building AI systems need to ask, as a routine part of their development process, who might try to attack this system, what they would be trying to achieve and what tools and knowledge they would have access to. Clean-data accuracy metrics, meanwhile, need to be supplemented or replaced by measures that actually test how a system behaves under adversarial pressure [32]. A system that achieves 98 percent accuracy on a held-out test set but fails catastrophically when a patient adversary probes its boundaries is not a secure system, whatever its benchmark score says. Continuous monitoring, scheduled red-team exercises and post-deployment auditing, all standard practice in mature cybersecurity operations should become equally standard in AI system management [33]. The policy and governance dimensions of this problem are equally urgent. Most existing AI regulatory frameworks focus on transparency, fairness and data protection. These are important concerns, but they leave adversarial robustness largely unaddressed [34]. This is a gap that regulators and standards bodies need to close, particularly for AI systems deployed in safety-critical contexts. Equally pressing is the question of accountability. When an AI system is compromised through a data poisoning attack, responsibility may fall on any number of parties , the organisation that collected the training data, the team that built the model, the platform that deployed it or the operator that failed to monitor it [35]. Without clear accountability frameworks, incidents are harder to investigate, remediate, and learn from. Finally, governance must grapple honestly with the tension between openness,which drives scientific progress and reproducibility and the security risks that come with publishing model architectures and releasing APIs [36]. Getting this balance right is genuinely difficult, but it is a challenge that cannot be deferred.

5. CONCLUSION

The picture that emerges from this meta-analysis is one that should concern anyone building, deploying or regulating AI systems. Adversarial attacks are not a fringe research concern or a laboratory curiosity, they are a structural feature of the threat landscape facing modern machine learning. The vulnerabilities that enable them are not random; they arise from deep properties of how learning-based systems work and they show up reliably across training, inference and deployment phases regardless of the specific architecture or application involved [23-25]. Narrowly targeted defences, designed to address one attack type in one setting are simply not adequate to this challenge. Perhaps the most important practical lesson from this analysis is the need to close the gap between how AI systems are evaluated and how they are actually used. Benchmark performance is not the same as operational resilience, and the field needs evaluation frameworks that reflect this distinction [26,27]. For engineers, this means embedding threat modelling into development workflows and treating adversarial testing as an ongoing operational responsibility rather than a one-time gate [31-33]. For policymakers, it means updating governance frameworks to explicitly address adversarial robustness alongside the fairness and transparency requirements that currently dominate AI regulation [34-36]. And for the research community,

DOI: 10.17148/IJIREEICE.2026.14614

it means building closer working relationships across AI, cybersecurity, governance and social science, because the challenge of securing AI systems at scale is not one that any single discipline can meet on its own.

REFERENCES

[1] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349(6245):255–260.
[2] Brundage M, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv. 2018;arXiv:1802.07228.
[3] Szegedy C, et al. Intriguing properties of neural networks. arXiv. 2014;arXiv:1312.6199.
[4] Papernot N, et al. The limitations of deep learning in adversarial settings. IEEE EuroS&P. 2016;372–387.
[5] Eykholt K, et al. Robust physical-world attacks on deep learning models. CVPR. 2018;1625–1634.
[6] Grosse K, et al. Adversarial examples for malware detection. ESORICS. 2017;62–79.
[7] Wallace E, et al. Universal adversarial triggers for attacking NLP models. EMNLP. 2019;2153–2162.
[8] Madry A, et al. Towards deep learning models resistant to adversarial attacks. ICLR. 2018.
[9] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security. ICML. 2018;274–283.
[10] Biggio B, Roli F. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition. 2018;84:317–331.
[11] Gilmer J, et al. Adversarial spheres. ICLR Workshop. 2018.
[12] Huang L, et al. Adversarial machine learning. ACM CCS. 2011;43–58.
[13] Tramèr F, et al. Ensemble adversarial training: Attacks and defenses. ICLR. 2018.
[14] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. IEEE S&P. 2017;39–57.
[15] Gu T, Dolan-Gavitt B, Garg S. BadNets: Vulnerabilities in the ML model supply chain. arXiv. 2017;arXiv:1708.06733.
[16] Papernot N, et al. SoK: Security and privacy in machine learning. IEEE EuroS&P. 2018;399–414.
[17] Kitchenham B, Charters S. Guidelines for systematic literature reviews in software engineering. EBSE. 2007.
[18] NIST. Adversarial machine learning: Taxonomy of attacks and mitigations. NIST IR 8269. 2020.
[19] Dong Y, et al. Benchmarking adversarial robustness on image classification. CVPR. 2020;321–331.
[20] Carlini N, et al. On evaluating adversarial robustness. arXiv. 2019;arXiv:1902.06705.
[21] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines. ICML. 2012.
[22] Zhang H, et al. Theoretically principled trade-off between robustness and accuracy. ICML. 2019.
[23] Szegedy C, et al. Intriguing properties of neural networks. Proc. ICLR. 2014.
[24] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. Proc. ICLR. 2015.
[25] Biggio B, Roli F. Wild patterns. Pattern Recognit. 84:317–331. 2018.
[26] Carlini N, Wagner D. Evaluating robustness of neural networks. IEEE S&P. 2017.
[27] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision. IEEE Access. 2018.
[28] Wallace E, et al. Universal adversarial triggers for NLP. EMNLP. 2019.
[29] Huang S, et al. Adversarial attacks on neural network policies. ICLR. 2017.
[30] Tramèr F, et al. Stealing machine learning models via prediction APIs. USENIX. 2016.
[31] Shostack A. Threat Modeling: Designing for Security. Wiley. 2014.
[32] Tsipras D, et al. Robustness may be at odds with accuracy. ICLR. 2019.
[33] Microsoft. The AI Red Teaming Handbook. Microsoft Security. 2023.
[34] European Commission. Proposal for a Regulation on Artificial Intelligence (AI Act). 2021.
[35] Floridi L, Cowls J. A unified framework of five principles for AI in society. Harv. Data Sci. Rev. 2019.
[36] Kearns M, Roth A. The Ethical Algorithm. Oxford University Press. 2019.
[37] NIST. AI Risk Management Framework (AI RMF 1.0). 2023.
[38] Croce F, et al. RobustBench: A standardized adversarial robustness benchmark. NeurIPS. 2021.