American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX • Website: www.ajamri.com • Email: editor@ajamri.com

Adversarial Robustness in Artificial Intelligence: Improving Reliability of Intelligent Systems in High-Stakes Environments

Dr. Ethan Alexander Cole

Associate Professor, University of Technology Sydney, Australia

Abstract

Artificial intelligence systems increasingly participate in decisions involving medical diagnosis, autonomous transportation, financial risk assessment, critical infrastructure, public safety, and cybersecurity. Although these systems can achieve high predictive accuracy under conventional testing conditions, carefully constructed adversarial inputs may cause substantial and sometimes imperceptible changes in their outputs. Such vulnerability raises a fundamental question: whether accuracy measured on clean data provides sufficient evidence of reliability when an intelligent system operates in a hostile, uncertain, or safety-critical environment. This study examines adversarial robustness as a multidimensional property involving resistance, detection, recovery, uncertainty management, operational monitoring, and governance.

A simulation-based comparative methodology was applied to five defense configurations: an undefended baseline model, input preprocessing, adversarial training, adversarial training combined with ensemble monitoring, and a certified hybrid defense. The configurations were evaluated through attack resistance, clean-data performance, attack detection, recovery capability, assurance strength, and operational feasibility. The resulting composite scores were 38, 54, 72, 76, and 84, respectively. These values are illustrative scenario outputs rather than empirical benchmark findings. The comparison indicates that isolated preprocessing offers only limited protection, while adversarial training produces a substantial improvement in resistance. The strongest simulated performance was obtained through a layered configuration integrating robust training, certified analysis, calibrated uncertainty, monitoring, and human-directed fallback procedures.

The findings suggest that adversarial robustness cannot be secured through a single defensive algorithm. High-stakes reliability requires protection throughout the artificial intelligence lifecycle, including data provenance, threat modeling, secure training, adaptive testing, model verification, runtime monitoring, incident response, and accountable human oversight. The study proposes that robustness claims should be conditional, attack-specific, and supported by transparent evaluation rather than expressed as universal guarantees. This integrated approach can improve the dependability of intelligent systems while acknowledging the technical and institutional limitations that remain.

Keywords: adversarial robustness; adversarial machine learning; reliable artificial intelligence; high-stakes systems; adversarial training; certified robustness; model assurance; intelligent-system security

1. Introduction

Artificial intelligence has progressed from a primarily analytical technology to an operational component of systems that directly affect human safety, access to services, financial opportunity, mobility, and infrastructure. Machine-learning models now assist clinicians in interpreting medical

images, support fraud detection, guide autonomous vehicles, identify cybersecurity threats, and prioritize emergency responses. In such settings, a prediction is not merely an informational output. It may influence a treatment decision, authorize a transaction, activate a mechanical response, or determine whether a threat receives immediate attention. Reliability therefore becomes a socio-technical requirement rather than a narrow measure of predictive performance.

Conventional model evaluation usually assumes that training and operational data are sufficiently similar and that errors arise naturally rather than through deliberate manipulation. Adversarial environments violate these assumptions. An attacker may modify an image, message, sensor signal, medical record, or network feature to produce a desired model response. The modification may be small enough to escape casual human inspection while being sufficient to alter the model's classification. Adversarial activity can occur during training through poisoned samples and manipulated labels, during inference through evasive inputs, or after deployment through model extraction, privacy attacks, and repeated probing.

Szegedy and colleagues demonstrated that neural networks could be induced to make incorrect predictions through carefully designed input perturbations. Goodfellow, Shlens, and Szegedy subsequently proposed the fast gradient sign method and connected adversarial vulnerability to the locally linear behavior of high-dimensional models. These findings challenged the belief that strong test-set accuracy necessarily indicated stable learned representations. Later research showed that adversarial weakness was not restricted to one architecture, dataset, or attack method. Decision trees, reinforcement-learning systems, speech models, biometric systems, and other learning configurations may also be vulnerable.

The implications become more serious in high-stakes environments because the costs of failure are asymmetric. A false-negative medical prediction may delay treatment, whereas a false-positive result may expose a patient to unnecessary intervention. In autonomous transport, a momentary perception error can affect physical safety. In financial services, manipulated decisions can create economic loss or discriminatory exclusion. A robustness assessment must consequently consider not only whether an attack succeeds, but also the severity, reversibility, detectability, and distribution of resulting harm.

This study examines how adversarial robustness can improve the reliability of intelligent systems. It compares selected defense configurations through a transparent simulation and develops a lifecycle-oriented interpretation of robustness. The central argument is that dependable artificial intelligence requires layered technical defenses combined with monitoring, recovery, documentation, and human authority. The paper does not claim completed experimental validation. Its numerical analysis is explicitly illustrative and intended to demonstrate an evaluation framework that can later be applied to verified datasets and operational systems.

2. Review of Literature

2.1 Development of Adversarial Attack Research

Early adversarial-machine-learning research established that a model could behave unpredictably when an attacker introduced purposefully optimized modifications. Szegedy and colleagues found that deep neural networks were vulnerable to perturbations that appeared negligible to human observers. Goodfellow, Shlens, and Szegedy explained part of this phenomenon through linear behavior in high-dimensional input spaces and developed an efficient gradient-based attack. Their work made adversarial-example generation computationally accessible and helped transform robustness into a major research problem.

Papernot and colleagues expanded the threat model by demonstrating transferability: an adversarial example designed against one model could sometimes mislead another. This finding made black-box attacks more practical because an attacker did not always require full access to the target

architecture or parameters. Carlini and Wagner later introduced optimization-based attacks capable of bypassing several proposed defenses. Their results reinforced the importance of evaluating defenses against adaptive adversaries who understand and deliberately respond to the protection mechanism.

Adversarial manipulation extends beyond norm-bounded image perturbations. Engstrom and colleagues showed that models can be vulnerable to spatial transformations such as translation and rotation. Poisoning attacks alter the training process through corrupted samples, while backdoor attacks create hidden behavior activated by a trigger. Model-inversion and membership-inference attacks target information confidentiality. The literature therefore indicates that adversarial robustness must account for attack objectives, attacker knowledge, access level, lifecycle stage, and permissible modification constraints.

2.2 Defensive Learning and Certified Protection

Adversarial training is among the most extensively examined empirical defenses. The method generates or approximates adversarial examples during training and optimizes the model to classify them correctly. Madry and colleagues framed this process as robust optimization:

$$\theta \min E(x,y)\sim D[\delta \in S \max L(f\theta(x+\delta),y)]$$

where θ denotes model parameters, D is the data distribution, L is the loss function, and S specifies the permitted perturbation set. The inner maximization identifies damaging perturbations, while the outer minimization attempts to reduce the resulting worst-case loss.

Adversarial training has produced substantial empirical improvements, but it has limitations. It increases computational cost, can reduce clean-data accuracy, and may specialize the model against attack families represented during training. A model trained against one perturbation norm or attack budget may remain vulnerable to another. Athalye, Carlini, and Wagner further demonstrated that some apparent defenses relied on gradient obfuscation rather than genuine resistance. When evaluated with appropriately adapted attacks, their protection weakened considerably.

Certified defenses attempt to provide mathematical guarantees under explicitly defined conditions. Wong and Kolter developed a convex outer approximation for certifying resistance against bounded perturbations. Cohen, Rosenfeld, and Kolter introduced randomized smoothing as a scalable method for constructing classifiers with certified resistance within an L2 radius. Certification is valuable because it does not depend solely on the failure of a particular attack implementation. Nevertheless, certified radii may be too limited for some practical threats, and guarantees ordinarily apply only to specified perturbation sets.

2.3 Robustness in High-Stakes Operational Systems

High-stakes deployment requires a wider conception of reliability than adversarial accuracy alone. A model may resist a bounded attack while failing under distributional change, sensor degradation, missing information, semantic manipulation, or implementation error. Conversely, a model that cannot classify every adversarial input may still operate safely if it recognizes uncertainty, declines to decide, activates an alternative sensor, or transfers authority to a qualified human operator.

Lütjens and colleagues examined certified adversarial robustness in deep reinforcement learning, highlighting its relevance to collision avoidance and other safety-sensitive applications. Their work demonstrates that a prediction should be assessed within the complete decision and control architecture. In autonomous systems, the severity of a perception error depends on redundancy, temporal consistency checks, control constraints, environmental conditions, and the availability of a safe fallback state.

The literature reveals a persistent gap between algorithmic robustness and operational assurance. Many studies report performance on standardized image datasets under narrowly specified attacks.

High-stakes systems, however, encounter physical constraints, strategic attackers, multiple data modalities, time pressure, and shifting environmental conditions. The present study addresses this gap by treating robustness as a composite of resistance, detection, recovery, assurance, and deployability.

3. Objectives of the Study

3.1 Analytical Objectives

The first objective is to clarify adversarial robustness as a multidimensional reliability requirement rather than a single benchmark score. This involves distinguishing resistance from detection, calibrated uncertainty, recovery, and formally bounded assurance. The study also seeks to explain why clean-data accuracy cannot independently establish the safety of a model exposed to deliberate manipulation.

The second objective is to compare five representative defense configurations through a common simulation-based framework. The comparison examines whether progressively layered protections provide stronger performance than isolated preprocessing or an undefended baseline model.

3.2 Operational and Governance Objectives

The study aims to connect algorithmic defense with the operational requirements of healthcare, finance, autonomous systems, and critical infrastructure. This includes the ability to recognize unusual inputs, defer uncertain decisions, maintain audit trails, preserve human control, and recover safely after a suspected attack.

A further objective is to formulate design principles for responsible robustness claims. Such claims should specify the attack model, perturbation limits, testing conditions, residual vulnerabilities, and boundaries of any formal guarantee. The study therefore treats transparency and governance as integral components of technical reliability.

4. Research Methodology

4.1 Research Design and Evidence Base

The study adopted a comparative simulation design informed by established adversarial-machine-learning literature. No human participants, operational organizations, confidential records, or live production models were involved. The analysis did not reproduce the numerical results of any individual published experiment. Instead, it developed illustrative values to demonstrate how a multilayer robustness assessment could be structured.

The evidence base included foundational and methodologically influential research on adversarial examples, transferability, adversarial training, optimization-based attacks, certified defenses, randomized smoothing, spatial attacks, gradient obfuscation, and robustness in reinforcement learning. Sources were selected for their direct relevance to adversarial reliability and their contribution to recognized attack or defense methodologies.

4.2 Defense Configurations

Five configurations were constructed. The baseline represented a conventionally trained model evaluated without a dedicated adversarial defense. The preprocessing configuration introduced input transformation and anomaly screening. The third configuration applied projected-gradient-based adversarial training. The fourth combined robust training with ensemble disagreement and runtime monitoring. The final configuration added certified analysis, uncertainty calibration, abstention rules, and human-directed fallback procedures.

These configurations should be understood as analytical scenarios rather than products. Their performance depends on model architecture, data quality, attack budget, task complexity,

implementation, and evaluation methodology. The simulated scores therefore describe comparative tendencies within the framework, not universal performance levels.

Table 1. Operational definition of the simulated defense configurations

Configuration	Principal protections	Primary limitation
Baseline model	Conventional training and clean-data validation	No dedicated adversarial resistance
Input preprocessing	Denoising, compression, transformation, and anomaly screening	Adaptive attacks may bypass transformations
Adversarial training	Training with optimized perturbations	Computational burden and attack-specific coverage
Training with ensemble monitoring	Robust training, model disagreement, and runtime alerts	Increased operational complexity
Certified hybrid defense	Robust training, certification, calibration, monitoring, abstention, and fallback	High implementation and verification cost

Note: Configurations and limitations are conceptual and literature-informed.

4.3 Measurement Framework and Scenario Construction

Six assessment dimensions were employed: attack resistance, clean-data utility, detection capability, recovery capacity, assurance strength, and operational feasibility. Each dimension was scored on a five-point ordinal scale.

The weighted composite robustness score was calculated as:

$$CRS_j = 100(0.25R_j + 0.15U_j + 0.15D_j + 0.15C_j + 0.20A_j + 0.10F_j)/5$$

where CRS_j represents the composite robustness score of configuration j; R represents attack resistance; U, clean-data utility; D, detection capability; C, recovery capacity; A, assurance strength; and F, operational feasibility.

The simulation assumed a high-stakes classification system exposed to a mixture of white-box and black-box evasion attempts. It also assumed that erroneous high-confidence outputs were more harmful than cautious abstention. Accordingly, certified bounds, calibrated uncertainty, monitoring, and safe fallback received explicit value within the assessment.

4.4 Analytical Procedure and Integrity Controls

The analysis proceeded through threat identification, configuration definition, criterion scoring, weighted aggregation, comparative interpretation, and sensitivity review. The robustness score was not treated as a statistical estimator. No inferential significance tests were applied because the underlying values were constructed scenarios rather than sampled observations.

Integrity controls included visible labeling of simulated values, separation of literature findings from author-developed analysis, avoidance of fabricated participants or organizations, and explicit recognition of limitations. A real empirical application would require version-controlled code,

predefined threat models, verified datasets, repeated attack runs, confidence intervals, attack-parameter disclosure, and independent replication.

5. Analysis

5.1 Comparative Robustness Results

The baseline model received a composite score of 38. This result represents a system that may retain acceptable utility on conventional test data but offers limited resistance, detection, recovery, or assurance when deliberately manipulated. The score does not imply that every undefended model will achieve the same value. It illustrates the inadequacy of treating ordinary predictive accuracy as evidence of adversarial reliability.

Input preprocessing increased the score to 54. Denoising, compression, feature squeezing, or transformation can remove some superficial perturbations and detect unusual inputs. However, a knowledgeable attacker may optimize against the preprocessing pipeline itself. This configuration therefore improves resistance only moderately unless its transformations are combined with adaptive testing and additional controls.

Its improvement arose primarily from stronger attack resistance and better alignment between training and worst-case evaluation. The training-plus-monitoring configuration achieved
Its incremental advantage reflected ensemble disagreement, anomaly monitoring, and improved capacity to identify uncertain inputs. The certified hybrid defense achieved the highest simulated score of 84 because it combined empirical resistance with bounded assurance, monitoring, calibrated abstention, and recovery procedures.

Table 2. Simulated adversarial-robustness assessment

Defense configuration	Attack resistance	Detection and recovery	Assurance strength	Composite score
Baseline model	1.8/5	1.5/5	1.0/5	38
Input preprocessing	2.6/5	2.8/5	1.7/5	54
Adversarial training	4.0/5	3.1/5	3.2/5	72
Training and ensemble monitoring	4.1/5	4.2/5	3.4/5	76
Certified hybrid defense	4.5/5	4.4/5	4.6/5	84

Source: Author-developed simulation informed by established adversarial-machine-learning literature.
Caution: These values are illustrative. They are neither measurements from a real system nor pooled estimates from published benchmarks.

5.2 Interpretation Across High-Stakes Domains

In medical artificial intelligence, adversarial resistance must coexist with clinically calibrated uncertainty. A defense that improves attack resistance while increasing false negatives could create unacceptable patient risk. Robustness testing should therefore be stratified by clinical condition, imaging

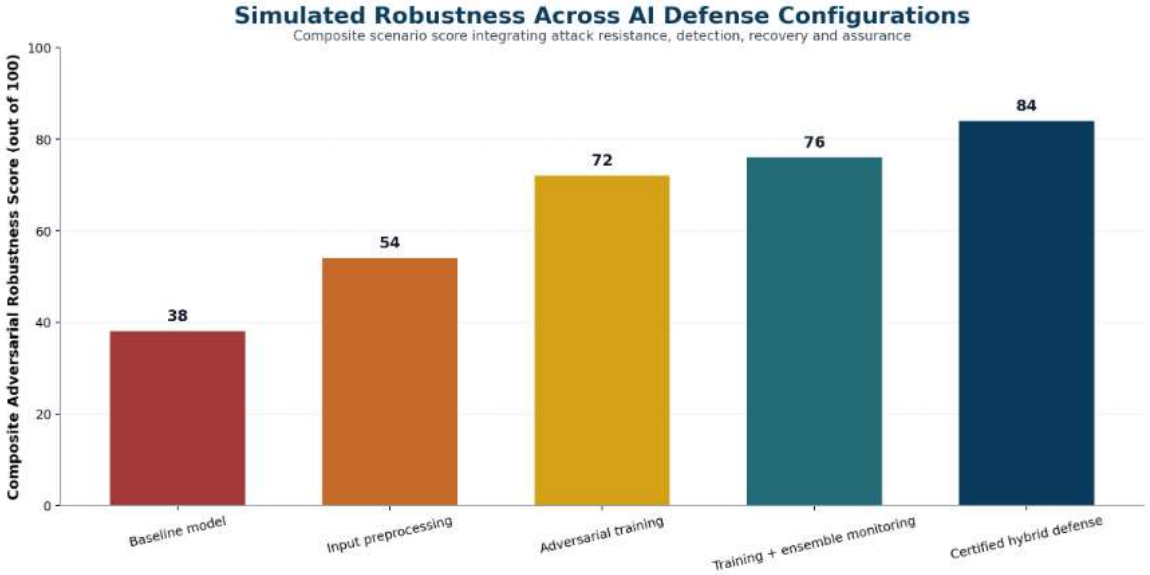
device, patient subgroup, and potential error severity. Suspicious or low-confidence cases should be referred to qualified clinicians rather than automatically rejected or classified.

In autonomous systems, temporal and physical consistency provide additional defensive signals. A traffic sign detected in one frame should remain spatially consistent across subsequent frames and agree with map, lidar, and vehicle-state information. Ensemble monitoring can therefore incorporate different sensors and models rather than merely combining multiple versions of the same classifier. A safe system should also reduce speed, increase following distance, or request human control when perception becomes unreliable.

Financial and critical-infrastructure systems require attention to strategic behavior. Attackers may repeatedly probe decision boundaries, manipulate transaction attributes, or poison historical records. Robustness must therefore include rate limiting, identity controls, provenance validation, logging, change detection, and post-incident investigation. Algorithmic defense without access control and cybersecurity integration would provide incomplete protection.

5.3 Robustness Graph

Graph 1. Simulated Robustness Across AI Defense Configurations



Supporting data: Baseline model, 38; input preprocessing, 54; adversarial training, 72; training with ensemble monitoring, 76; and certified hybrid defense, 84.

Interpretation: The simulated comparison shows that preprocessing alone produces limited improvement. Adversarial training substantially strengthens the model, while a layered system combining training, monitoring, certification, uncertainty management, and fallback procedures achieves the highest composite score.

Key finding: The 46-point difference between the baseline and certified hybrid configurations indicates that operational reliability depends on coordinated protection across multiple system layers rather than on a single defensive technique.

Source: Author-developed simulation. Values are illustrative and are not empirical benchmark results.

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

6. Challenges

6.1 Adaptive Attackers and Evaluation Limitations

Adversarial defense is difficult because attackers can adapt. A protection mechanism that succeeds against a fixed attack may fail when the adversary obtains information about its structure. Evaluation must therefore assume that the attacker understands the defense unless secrecy is itself an explicitly justified security control. Gradient-based, gradient-free, transfer, query-based, and physical attacks should be considered where relevant.

Attack success is also highly sensitive to perturbation budget, number of optimization steps, restart count, model access, preprocessing, and success criteria. Reporting a single adversarial-accuracy value without these details provides weak evidence. Robustness evaluation should disclose clean accuracy, robust accuracy, attack parameters, confidence intervals, failure cases, and computational cost.

6.2 Accuracy, Robustness, and Resource Trade-Offs

Robust optimization can create a tension between clean-data performance and adversarial resistance. The extent of this tension depends on task structure, model capacity, training procedure, and perturbation definition. In high-stakes settings, a small reduction in conventional accuracy may be acceptable if the system gains substantial resistance and better-calibrated uncertainty. However, this trade-off must be evaluated against domain-specific harm rather than assumed to be universally beneficial.

Adversarial training can require repeated inner-loop optimization, increasing energy use and computational expense. Certified methods may impose further scalability constraints. Smaller organizations and public institutions may therefore lack the infrastructure required to reproduce advanced robustness procedures. Shared evaluation resources, reproducible benchmarks, and transparent reference implementations could reduce this disparity.

6.3 Governance and Human-Factor Challenges

Technical robustness can be undermined by weak operational governance. An organization may deploy a robustly trained model but fail to monitor data drift, record model changes, respond to alerts, or establish clear authority for shutdown and fallback decisions. Responsibility may also be fragmented among developers, vendors, cybersecurity teams, operators, and senior management.

Human oversight is not automatically effective. Operators may experience automation bias, alert fatigue, insufficient training, or unclear accountability. A reliable fallback process must specify who receives an alert, what evidence is displayed, how quickly intervention is required, and what happens when human review is unavailable. Robustness must therefore be embedded in an institutional procedure rather than treated solely as a model property.

7. Discussion

7.1 Adversarial Robustness as a System-Level Property

The findings indicate that model-level resistance is necessary but insufficient. A classifier may resist a defined perturbation while the surrounding system remains vulnerable through insecure data pipelines, manipulated sensors, weak authentication, or poorly designed operator interfaces. Conversely, a model with limited standalone robustness may be used more safely when uncertainty estimation, redundant evidence, monitoring, and human review constrain its authority.

This interpretation changes the design question. Instead of asking whether a model is adversarially robust in a universal sense, developers should ask which attacks are plausible, which

failures are tolerable, which guarantees apply, and how the system detects and contains unanticipated behavior. Such questions produce conditional and testable assurance claims.

The certified hybrid configuration performed most favorably because it represented defense in depth. Robust training reduced susceptibility, certification supplied bounded guarantees, monitoring identified abnormal behavior, calibration supported meaningful confidence estimates, and fallback procedures limited the consequences of unresolved uncertainty. The result does not establish that every hybrid system will be superior. It demonstrates why complementary controls can address different failure modes.

7.2 Reliability in High-Stakes Decision Processes

Reliability should be evaluated in relation to the complete decision process. In healthcare, a model's output may be advisory, confirmatory, or determinative. The required robustness level should increase with its decision authority and the severity of potential harm. The same principle applies to financial lending, industrial control, and autonomous navigation.

A useful assurance framework should combine adversarial accuracy with selective risk. A model that abstains on suspicious inputs may have lower coverage but produce safer accepted predictions. Relevant measures include robust error, false-acceptance rate, abstention rate, detection latency, recovery time, and severity-weighted harm. This wider measurement system is more informative than a single accuracy figure.

Reliability also requires continuous reassessment. Attack methods evolve, software dependencies change, sensors degrade, and operational data shift. A robustness evaluation performed before initial deployment cannot establish indefinite safety. Periodic red-team testing, post-deployment monitoring, incident analysis, and controlled model updates should form part of the maintenance lifecycle.

7.3 Implications for Research and Practice

Researchers should evaluate defenses against adaptive and diverse attacks, report complete configurations, and distinguish empirical resistance from formal certification. Negative results and failed defenses should be documented because they reveal where apparently strong protections rely on incomplete evaluation. Benchmark performance should be supplemented with physical, semantic, and domain-valid attacks.

Developers should maintain a threat model that identifies protected assets, attacker capabilities, attack surfaces, consequences, and accepted residual risks. Data pipelines should support provenance checks and integrity controls. Models should be tested under clean, corrupted, shifted, and adversarial conditions. Deployment should include monitoring thresholds, escalation procedures, rollback capability, and audit records.

Organizations operating high-stakes systems should require evidence proportionate to the possible harm. A low-impact recommendation engine may justify empirical stress testing, whereas an autonomous safety function may require stronger certification, redundancy, and independent validation. Procurement contracts should specify robustness testing, vulnerability disclosure, update responsibilities, and access to relevant technical documentation.

8. Conclusion

Adversarial robustness has become an essential condition for the responsible use of artificial intelligence in high-stakes environments. Models that perform well on conventional test data may remain vulnerable to deliberate manipulation, unexpected transformations, poisoned data, or strategic probing. Clean accuracy therefore cannot serve as an adequate proxy for operational reliability.

The simulation-based comparison demonstrated a progressive improvement from an undefended baseline to a certified hybrid defense. Input preprocessing offered modest protection, while adversarial training produced a stronger improvement. Combining robust training with ensemble monitoring increased detection and recovery capacity. The highest simulated score was obtained when these measures were integrated with certified analysis, calibrated uncertainty, abstention, and human-directed fallback.

The study's principal limitation is that its numerical results are illustrative rather than empirical. They should not be interpreted as performance estimates for any particular model or domain. Future research should apply the framework to verified datasets, clearly specified threat models, repeated adaptive attacks, physical-world conditions, and independent replication. Domain experts should also contribute to defining valid perturbations and severity-weighted outcomes.

The central conclusion is that adversarial robustness should be developed as a lifecycle-wide assurance capability. Secure data governance, robust optimization, certification, monitoring, recovery, documentation, and accountable human oversight must operate together. Intelligent systems will become more dependable not when every possible attack is assumed to have been eliminated, but when vulnerabilities are systematically tested, uncertainty is responsibly managed, and failures can be detected and contained before they produce serious harm.

References

1. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 274–283).
2. Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning* (pp. 1467–1474).
3. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84*, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023
4. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy* (pp. 39–57). https://doi.org/10.1109/SP.2017.49
5. Chen, H., Zhang, H., Boning, D., & Hsieh, C.-J. (2019). Robust decision trees against adversarial examples. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 1122–1131).
6. Cissé, M., Bojanowski, P., Grave, E., Dauphin, Y., & Usunier, N. (2017). Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 854–863).
7. Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 1310–1320).
8. Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., & Madry, A. (2019). Exploring the landscape of spatial robustness. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 1802–1811).
9. Gilmer, J., Ford, N., Carlini, N., & Cubuk, E. (2019). Adversarial examples are a natural consequence of test error in noise. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 2280–2289).
10. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.
11. Gu, S., & Rigazio, L. (2015). Towards deep neural network architectures robust to adversarial examples. In *International Conference on Learning Representations Workshop*.

12. Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. In *International Conference on Learning Representations Workshop*.
13. Lütjens, B., Everett, M., & How, J. P. (2020). Certified adversarial robustness for deep reinforcement learning. In *Proceedings of the Conference on Robot Learning* (pp. 1328–1337).
14. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.
15. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy* (pp. 372–387). https://doi.org/10.1109/EuroSP.2016.36
16. Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. *arXiv*. https://doi.org/10.48550/arXiv.1605.07277
17. Raghunathan, A., Steinhardt, J., & Liang, P. (2018). Certified defenses against adversarial examples. In *International Conference on Learning Representations*.
18. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.
19. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*.
20. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2019). Robustness may be at odds with accuracy. In *International Conference on Learning Representations*.
21. Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., & Gu, Q. (2019). On the convergence and robustness of adversarial training. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6586–6595).
22. Wong, E., & Kolter, J. Z. (2018). Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 5286–5295).
23. Zhang, H., Yu, Y., Jiao, J., Xing, E. P., El Ghaoui, L., & Jordan, M. I. (2019). Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 7472–7482).