SECURITY FOR AI [*]

Alvaro Velasquez
Information Directorate
Air Force Research Laboratory
Rome, NY 13441
alvaro.velasquez.1@us.af.mil

Rickard Ewetz
ECE Department
University of Central Florida
Orlando, FL 32765
rickard.ewetz@ucf.edu

Amro Awad
ECE Department
North Carolina State University
Raleigh, NC 27695
amro.awad@ncsu.edu

Sumit Kumar Jha
Computer Science Department
University of Texas at San Antonio
San Antonio, TX 78255
sumit.jha@utsa.edu

ABSTRACT

As AI systems become pervasive in our daily lives, they may create new opportunities for security attacks that are hitherto unobserved in classical software and hardware systems. In this report, we briefly sketch recent advances and opportunities in the security of AI systems. Special emphasis has been placed on technology gaps in this area, including missing theoretical foundations, desirable algorithmic advances, and application areas of critical interest.

Keywords Security · AI · Adversarial Robustness · Certified Robustness · Black-box Attacks · Patch Attacks · Natural Perturbations · Integrated Gradients · Class Activation Maps · Neural SDE · Normalizing Flows · Attack Models

1 Introduction

The problem of securing Artificial Intelligence (AI) is increasingly prevalent as the adoption of these systems is expected in the near future across a vast variety of domains. The consequences of failing to secure such systems is evident in the autonomous driving sector, where faults in the logic and assumptions of the driving agent have caused accidents during traffic merging (Ziegler 2016). More significantly and to tragic effect was the fatality caused by the inferencing module of a Tesla autopilot that failed to recognize a white truck against the bright background sky (Lambert 2016). In order to address the plethora of security concerns implicit in such failures, it is important to consider the security of AI across its many stages of operation, including the data collection, architecture design, training time, test time, and inference time phases. Indeed, the challenges and opportunities for securing AI systems differ across each of the foregoing. For example, at data collection, there is a risk of data poisoning attacks whereby labels in an existing dataset can be manipulated to reflect erroneous relationships between input data and the predicted output class (Schwarzschild et al. 2021). Such attacks can me mitigated to some extent by leveraging techniques from data sanitization (Steinhardt, Koh, and P. S. Liang 2017).

It is also important to leverage and adapt the fundamental results established in conventional security domains, such as traditional software testing, where the notion of code coverage as a metric for deriving effective tests has found powerful analogues for AI systems in the form of neuron coverage. The remainder of this report is organized as follows. A brief overview of digital attacks and their corresponding defenses is presented in Section 2. This is followed in Section 3 with a similar treatment of attacks and defenses in the physical domain. Section 4 outlines approaches in the area of explainable AI that facilitate the interpretation of AI dynamics and outcomes to enable the human end-user to address security concerns. We conclude and propose future avenues of inquiry for secure AI systems in Section 5.

[*] *Thanks to the National Science Foundation Secure and Trustworthy Computing (SaTC) Program for supporting the breakout session at the 2022 NSF SaTC Meeting. This report summarizes a few of the topics on Security of AI discussed at this meeting. The views and opinions expressed in this article are those of the authors and do not reflect the views of the U.S. government.*

2 Digital Adversarial Attacks and Defenses

Adversarial attacks on modern deep learning systems (I. J. Goodfellow, Shlens, and Szegedy 2014) and more classical machine learning systems (Ramanathan et al. 2016) have been well studied for nearly a decade now. Suitably-crafted small and often imperceptible changes to the input cause AI systems to produce different responses on otherwise similar inputs.

2.1 Attacks on AI Systems

There has been a plethora of digital attacks on AI systems and we make no attempt to create a compendium of the same. Instead, we highlight a few attacks and then discuss the ongoing work on defending against such digital attacks by training models that are inherently robust.

2.1.1 Gradient-based Attacks

Popular attacks, such as Fast Gradient Sign Method (I. J. Goodfellow, Shlens, and Szegedy 2014), Basic Iterative Method (Kurakin, I. J. Goodfellow, and Bengio 2018), Carlini-Wagner (Carlini and Wagner 2017), and the Projected Gradient Descent, have been deployed in a number of open-source tools, such as Cleverhans (Papernot et al. 2016), the IBM Adversarial Robustness Toolbox (Nicolae et al. 2018) and TorchAttacks (Kim 2020) . These attacks rely on the back-propagation of gradients and the optimization engine to identify inputs that cause the model to change its prediction.

2.1.2 Black-box Attacks

Neural network models deployed as black-box models with no explicit access to the parameters of the model have been attacked (P.-Y. Chen et al. 2017; Cheng et al. 2018) by explicitly querying the model on multiple parameters, zeroth-order optimization (J. Chen, Jordan, and Wainwright 2020), or distilling a surrogate model possibly using generators J. Zhang et al. 2022 and exploiting the transferability of adversarial attacks. In one foundational approach (Brendel, Rauber, and Bethge 2017), a first adversarial example is obtained by relaxing the norms, and then the norm of the adversarial example is optimized sequentially.

2.2 Defenses against Digital Adversarial Attacks

Several defenses against digital adversarial attacks have been suggested with varying degrees of success. A few defenses published in elite venues have been broken by carefully crafted attacks, and certified defenses have allowed one to reason mathematically about the quality of a proposed defense.

2.2.1 Adversarial Training

Adversarial training (Madry et al. 2017) uses counterexamples generated by attacks and re-trains the model to be correct on these counterexamples. This is similar to the use of counterexamples to refine abstract models (Clarke et al. 2000; Sumit K Jha et al. 2007) during the static analysis of software as well as hardware systems. Several clever algorithms (Wong, Rice, and J. Z. Kolter 2020; Shafahi et al. 2019) that advance adversarial training neural networks have been proposed.

2.2.2 Certified Defenses

Several algorithms for certifying the robustness of neural networks have been proposed (Raghunathan, Steinhardt, and P. Liang 2018; Raghunathan, Steinhardt, and P. S. Liang 2018; B. Li et al. 2019). Randomized smoothing has been investigated and an algorithm for certifying the defense has been popularly used (Cohen, Rosenfeld, and Z. Kolter 2019).

2.3 Future Work

Several interesting directions for future research remain open:

1. There is a need to develop metrics for measuring the robustness of AI models by using metrics aligned with human perception (L. Zhang et al. 2011; Xue et al. 2013), instead of Euclidean and other purely mathematical norms.

2. While a lot of effort has been put into robustness measured by norms, there needs to be a deeper investigation into the query complexity (Gluch and Urbanke 2021; D. Lee et al. 2022) of adversarial attacks for different models and associated defenses.
3. AutoAttack and similar tools (Croce and Hein 2020; Croce, Andriushchenko, et al. 2020) have made it much easier to evaluate the robustness of models without the need to choose multiple attack parameters explicitly. However, human attackers have often performed much better than automated attack algorithms and there may be a need to focus on better automation of adversarial attacks as a preliminary evaluation of defenses.
4. The rapid creation of synthetic data (Ramesh et al. 2022; Nichol et al. 2021) allows us to re-visit the defense of AI models against adversarial attacks. It is likely that models, such as DALL-E 2, trained on large synthetic data or suitably-crafted synthetic data can provide a new impetus to the design of robust AI models.

3 Physical Adversarial Attacks and Defenses

Physical adversarial attacks refer to attacks that can be applied in the real world, where the input data to an AI system is collected from sensors such as cameras, microphones, etc. This is in contrast to digital attacks where the adversary has direct access to the input of the AI system. Different forms of physical adversarial attacks have been investigated (T. B. Brown et al. 2017; Kurakin, I. Goodfellow, and Bengio 2016). Adversarial objects that can be 3D-printed were demonstrated in Athalye et al. (2017). The use of glasses that can fool facial recognition software has been illustrated (Sharif et al. 2016). A practical approach may be to create adversarial patches and insert them into the real world in the form of a sticker (Eykholt et al. 2017).

3.1 Patch Attacks on AI Systems

Patch attacks cause misclasification by introducing a perceptible but spatially constrained change to an image. An adversarial patch that can be inserted anywhere on an image and cause misclassification to a target class was proposed in T. B. Brown et al. (2017). Adversarial patches are constructed by a grey patch on multiple input images while applying transformations such as translation, scaling, and rotation. Next, the pixels of the patch are updated using iterative gradient decent to maximize misclassification into a target class. Adversarial patches covering only 2% of the input image were shown to cause misclassification (Karmon, Zoran, and Goldberg 2018) while not being placed on top of the object in the input image. The use of texture based patch attacks have been investigated (Fernandes and Sumit Kumar Jha 2020; Yang et al. 2020). The texture and the placement location of the patch are often optimized using reinforcement learning.

With the rise of vision transformers (ViTs) for image classification, it was speculated that ViTs had inherent robustness to various attacks. In Fu et al. (2022), it was demonstrated that even ViTs are vulnerable to patch attacks. Recently, to circumvent patch detection, patch attacks that mimic objects in the natural environment were proposed (Hu et al. 2021).

3.2 Attacks using Natural Perturbations

Another interesting class of attacks is based on inserting natural perturbations like fog, rain, haze, snow, and other phenomena into the inputs of AI systems (Ozdag et al. 2019). This idea of injecting fake weather has recently been further developed and implemented in the fakeWeather system Marchisio et al. 2022. There is a need to investigate more physically realizable attacks using natural perturbations Gao et al. 2021.

3.3 Defences Against Physical Adversarial Attacks

Several successful efforts (Xiang et al. 2021; McCoyd et al. 2020) have been established for defending against patch attacks. In one methodology, the activations of the neural networks are clipped in order to ensure that a small area of the input cannot overpower the rest of the input (Yu et al. 2021). Other recent efforts have included certified defenses against patch attacks (Metzen and Yatsura 2021).

3.4 Future work

Several interesting directions for future research remain open:

1. There is a need to develop automated frameworks for the evaluation of the effectiveness of adversarial patch attacks and defences. While the APRICOT (Braunegg et al. 2019) and CARLA (Nesti et al. 2022) frameworks provide a promising starting point, there is an urgent need to develop efficient adaptive attacks.

2. Patch attacks often create bright patches that are readily identified by the human visual system. Algorithms for creating more subtle patches and defenses against the same remain an area of active interest.
3. Natural perturbations like fake weather (Ozdag et al. 2019) require novel metrics for measuring model robustness to such real-world perturbations. Defenses that focus on such natural perturbations need special attention for the safety of AI systems employed in autonomous applications like self-driving cars. Defenses for such attacks using graphs (Acharya et al. 2022) and manifolds (Jang, Susmit Jha, and Somesh Jha 2020) need to be further investigated.

4 Explainable Artificial Intelligence

Since AI models may continue to be susceptible to adversarial attacks, there is great value in explaining AI decisions to human experts. This will enable the detection of attacks on high-assurance systems, and a better understanding of adversarial defenses on the learning dynamics of AI models.

4.1 Enhanced Explanation Algorithms

There has been sustained work on explaining AI decisions for planning, rule-based and expert systems over the last few decades (Simmons 1988; Hammond 1990; Swartout 1983; Lane et al. 2005; Core et al. 2006). The rise of neural networks and their susceptibility to attacks has led to a renewed push to better explain the reasoning of neural networks Lundberg and S.-I. Lee 2017; Shrikumar, Greenside, and Kundaje 2017.

4.1.1 Class Activation Maps

In order to understand the influence of an input component x on a neural network F, the gradient $\frac{\delta F}{\delta x}$ serves as an effective first approximation. The higher the value of the derivative, the more the input influences the neural network. This idea combined with focusing on specific logits of the neural network and back-propagation through the neural network has given rise to saliency maps (Simonyan, Vedaldi, and Zisserman 2013) and variants of class activation maps (Selvaraju et al. 2017; Chattopadhay et al. 2018; Ramaswamy et al. 2020).

4.1.2 Integrated Gradients

Since the training of neural networks often leads to vanishing gradients with respect to the inputs, a path integral in the input space from a baseline input to the current input has been suggested as an approach for explaining neural networks. Such integrated gradients attributions have been shown to produce strong axiomatic properties (Sundararajan, Taly, and Yan 2017). A number of follow-up methods have been suggested for improving integrated gradients (Kapishnikov et al. 2021; Miglani et al. 2020; Pan, X. Li, and Zhu 2021) by focusing on saturation effects and choice of baseline inputs.

4.2 Designing Models for Better Explanations

A more recent phenomenon is the creation and training of AI models with explainability as a design goal. Security of AI systems may be enhanced and better subject to human audits if their decisions can be better explained.

4.2.1 Adversarial Robust Models

It has been shown that adversarial training leads to more concise explanations of deep neural networks (Chalasani et al. 2020). The relationship between better explanations and robust models has been studied deeply (Datta et al. 2021) for both in-distribution and out-of-distribution data (Augustin, Meinke, and Hein 2020). Adversarial models often have lower benign accuracy than traditional models which makes the deployment of such explainable models for security and explainability reasons less appealing to end users.

4.2.2 Homogeneous and Stochastic Models

The design of models and training methodologies for making more explainable models without directly focusing on adversarial robustness has also been studied. Inspired by SmoothGrad (Smilkov et al. 2017), neural SDE models have been shown to create more robust attributions (Sumit Jha, Ewetz, Velasquez, Ramanathan, et al. 2022; Sumit Jha, Ewetz, Velasquez, and Susmit Jha 2021). Non-negatively homogeneous deep neural networks (Hesse, Schaub-Meyer, and Roth 2021) have been shown to be structurally suitable for computing integrated gradients as the gradients scale linearly with linear scaling of the inputs.

4.3 Future Work

1. It has been shown that classical explanations are not well aligned with human intuition (Adebayo et al. 2018). Creating new methods that can robustly explain AI decisions in different contexts, such as graphs and DeepRL, may be needed.
2. The DALL-E 2 and similar systems (Ramesh et al. 2022; Nichol et al. 2021) enable the creation of both inputs and noise models of inputs that can provide novel data sets for training multi-modal explainability models.
3. While there has been some early work on designing better explainable models using non-negatively homogeneous models, the space of better explainable models needs to be investigated further. Variants of normalizing flows may be suited for designing models with enhanced explainability (Cunningham, Cobb, and Susmit Jha 2022).

5 Conclusions and Future Opportunities

5.1 DALL-E 2 and other data generation algorithms

Data augmentation algorithms have traditionally used simple manipulations, such as rotation and cropping of images. The rise of data synthesis algorithms, using neural SDEs and other diffusion approaches, gives a new opportunity to design algorithms for designing adversarially robust and secure AI models with better explainability (Ramesh et al. 2022; Nichol et al. 2021). Such ideas could include additional data, adversarial inputs and semantically rich data augmentation.

5.2 Neuro Symbolic Reasoning

There has been recent progress (Velasquez et al. 2021; Velasquez 2019; Brafman and De Giacomo 2019) in employing classical symbolic reasoning based on logic and automata to design new AI systems. The use of symbolic algorithms for reasoning about neural networks and the employment of neural networks for reasoning about symbolic systems require renewed attention for the design of secure AI systems. It should be highlighted that symbolic methods have been used for security applications in other software and hardware systems.

5.3 Security Beyond Euclidean Norms

An overwhelming fraction of research into AI security, including adversarial robustness and certified robustness, has focused on Euclidean and other norms that are easier to manipulate mathematically but have no effective basis in terms of measuring distances between inputs. The design of secure AI systems for practical deployments must prove robustness using metrics arising in the domain of application, such as visual-question answering (Walmer et al. 2022). For example, the distance between images may be captured by measures of human visual perception metrics (Setiadi 2021; Fan et al. 2019).

5.4 AI vs. Other Software and Hardware

Several attacks and defenses used in software and hardware systems have also been employed for AI applications with suitable modifications and justifications. There is a need to clearly delineate the threat model so as to focus clearly on the security of AI systems that exploit the additional attack surface created by data-driven deep learning and other AI systems.

5.5 New Applications: Social Networks and Malware Analysis

Securing AI systems used in social networks and malware analysis can have a profound effect on the security of our society. AI-assisted analysis of malware is a growing area of investigation (Abdelsalam, Gupta, and Mittal 2021; Or-Meir et al. 2019; Stamp, Alazab, and Shalaginov 2021); however, there is a need for a more open research cyber-infrastructure that brings together experts in AI and malware analysis.

Recent foundational work has been performed on the analysis of social networks (Oh and Kumar 2022; Yue et al. 2021; Anelli et al. 2022), with a focus on the robustness of graphs (Freitas et al. 2022). Metrics for adversarial robustness and similarity in AI systems used for analyzing social networks remain areas with great potential.

References

Swartout, William R (1983). "XPLAIN: A system for creating and explaining expert consulting programs". In: *Artificial intelligence* 21.3, pp. 285–325.

Simmons, Reid G (1988). "A Theory of Debugging Plans and Interpretations." In: *AAAI*, pp. 94–99.

Hammond, Kristian J (1990). "Explaining and repairing plans that fail". In: *Artificial intelligence* 45.1-2, pp. 173–228.

Clarke, Edmund et al. (2000). "Counterexample-guided abstraction refinement". In: *International Conference on Computer Aided Verification*. Springer, pp. 154–169.

Lane, H Chad et al. (2005). *Explainable artificial intelligence for training and tutoring*. Tech. rep. University of Southern California.

Core, Mark G et al. (2006). "Building explainable artificial intelligence systems". In: *AAAI*, pp. 1766–1773.

Jha, Sumit K et al. (2007). "Reachability for linear hybrid automata using iterative relaxation abstraction". In: *International Workshop on Hybrid Systems: Computation and Control*. Springer, pp. 287–300.

Zhang, Lin et al. (2011). "FSIM: A feature similarity index for image quality assessment". In: *IEEE transactions on Image Processing* 20.8, pp. 2378–2386.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2013). "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034*.

Xue, Wufeng et al. (2013). "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index". In: *IEEE transactions on image processing* 23.2, pp. 684–695.

Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy (2014). "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572*.

Kurakin, Alexey, Ian Goodfellow, and Samy Bengio (2016). *Adversarial examples in the physical world*. DOI: `10.48550/ARXIV.1607.02533`. URL: `https://arxiv.org/abs/1607.02533`.

Lambert, Fred (2016). "Understanding the fatal tesla accident on autopilot and the nhtsa probe". In: *Electrek, July* 1.

Papernot, Nicolas et al. (2016). "Technical report on the cleverhans v2. 1.0 adversarial examples library". In: *arXiv preprint arXiv:1610.00768*.

Ramanathan, Arvind et al. (2016). "Integrating symbolic and statistical methods for testing intelligent systems: Applications to machine learning and computer vision". In: *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, pp. 786–791.

Sharif, Mahmood et al. (2016). "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. Vienna, Austria: Association for Computing Machinery, pp. 1528–1540. ISBN: 9781450341394. DOI: `10.1145/2976749.2978392`. URL: `https://doi.org/10.1145/2976749.2978392`.

Ziegler, Chris (2016). "A google self-driving car caused a crash for the first time". In: *The Verge*.

Athalye, Anish et al. (2017). *Synthesizing Robust Adversarial Examples*. DOI: `10.48550/ARXIV.1707.07397`. URL: `https://arxiv.org/abs/1707.07397`.

Brendel, Wieland, Jonas Rauber, and Matthias Bethge (2017). "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models". In: *arXiv preprint arXiv:1712.04248*.

Brown, Tom B. et al. (2017). *Adversarial Patch*. DOI: `10.48550/ARXIV.1712.09665`. URL: `https://arxiv.org/abs/1712.09665`.

Carlini, Nicholas and David Wagner (2017). "Towards evaluating the robustness of neural networks". In: *2017 ieee symposium on security and privacy (sp)*. Ieee, pp. 39–57.

Chen, Pin-Yu et al. (2017). "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models". In: *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26.

Eykholt, Kevin et al. (2017). *Robust Physical-World Attacks on Deep Learning Models*. DOI: `10.48550/ARXIV.1707.08945`. URL: `https://arxiv.org/abs/1707.08945`.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems*, pp. 4765–4774.

Madry, Aleksander et al. (2017). "Towards deep learning models resistant to adversarial attacks". In: *arXiv preprint arXiv:1706.06083*.

Selvaraju, Ramprasaath R et al. (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017). "Learning important features through propagating activation differences". In: *arXiv preprint arXiv:1704.02685*.

Smilkov, Daniel et al. (2017). "Smoothgrad: removing noise by adding noise". In: *arXiv preprint arXiv:1706.03825*.

Steinhardt, Jacob, Pang Wei W Koh, and Percy S Liang (2017). "Certified defenses for data poisoning attacks". In: *Advances in neural information processing systems* 30.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). “Axiomatic attribution for deep networks”. In: *ICML*. JMLR. org, pp. 3319–3328.

Adebayo, Julius et al. (2018). “Sanity checks for saliency maps”. In: *Advances in neural information processing systems* 31.

Chattopadhay, Aditya et al. (2018). “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks”. In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, pp. 839–847.

Cheng, Minhao et al. (2018). “Query-efficient hard-label black-box attack: An optimization-based approach”. In: *arXiv preprint arXiv:1807.04457*.

Karmon, Danny, Daniel Zoran, and Yoav Goldberg (2018). *LaVAN: Localized and Visible Adversarial Noise*. DOI: 10.48550/ARXIV.1801.02608. URL: https://arxiv.org/abs/1801.02608.

Kurakin, Alexey, Ian J Goodfellow, and Samy Bengio (2018). “Adversarial examples in the physical world”. In: *Artificial intelligence safety and security*. Chapman and Hall/CRC, pp. 99–112.

Nicolae, Maria-Irina et al. (2018). “Adversarial Robustness Toolbox v1. 0.0”. In: *arXiv preprint arXiv:1807.01069*.

Raghunathan, Aditi, Jacob Steinhardt, and Percy Liang (2018). “Certified defenses against adversarial examples”. In: *arXiv preprint arXiv:1801.09344*.

Raghunathan, Aditi, Jacob Steinhardt, and Percy S Liang (2018). “Semidefinite relaxations for certifying robustness to adversarial examples”. In: *Advances in Neural Information Processing Systems* 31.

Brafman, Ronen I and Giuseppe De Giacomo (2019). “Planning for LTLf/LDLf Goals in Non-Markovian Fully Observable Nondeterministic Domains.” In: *IJCAI*, pp. 1602–1608.

Braunegg, A. et al. (2019). *APRICOT: A Dataset of Physical Adversarial Attacks on Object Detection*. DOI: 10.48550/ARXIV.1912.08166. URL: https://arxiv.org/abs/1912.08166.

Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter (2019). “Certified adversarial robustness via randomized smoothing”. In: *International Conference on Machine Learning*. PMLR, pp. 1310–1320.

Fan, Deng-Ping et al. (2019). “Scoot: A perceptual metric for facial sketches”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5612–5622.

Li, Bai et al. (2019). “Certified adversarial robustness with additive noise”. In: *Advances in neural information processing systems* 32.

Or-Meir, Ori et al. (2019). “Dynamic malware analysis in the modern era—A state of the art survey”. In: *ACM Computing Surveys (CSUR)* 52.5, pp. 1–48.

Ozdag, Mesut et al. (2019). “On the susceptibility of deep neural networks to natural perturbations”. In: *AISafety@ IJCAI*.

Shafahi, Ali et al. (2019). “Adversarial training for free!” In: *Advances in Neural Information Processing Systems* 32.

Velasquez, Alvaro (2019). “Steady-State Policy Synthesis for Verifiable Control”. In: *IJCAI*.

Augustin, Maximilian, Alexander Meinke, and Matthias Hein (2020). “Adversarial robustness on in-and out-distribution improves explainability”. In: *European Conference on Computer Vision*. Springer, pp. 228–245.

Chalasani, Prasad et al. (2020). “Concise explanations of neural networks using adversarial training”. In: *International Conference on Machine Learning*. PMLR, pp. 1383–1391.

Chen, Jianbo, Michael I Jordan, and Martin J Wainwright (2020). “Hopskipjumpattack: A query-efficient decision-based attack”. In: *2020 ieee symposium on security and privacy (sp)*. IEEE, pp. 1277–1294.

Croce, Francesco, Maksym Andriushchenko, et al. (2020). “Robustbench: a standardized adversarial robustness benchmark”. In: *arXiv preprint arXiv:2010.09670*.

Croce, Francesco and Matthias Hein (2020). “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks”. In: *International conference on machine learning*. PMLR, pp. 2206–2216.

Fernandes, Steven Lawrence and Sumit Kumar Jha (2020). “Adversarial attack on deepfake detection using RL based texture patches”. In: *European Conference on Computer Vision*. Springer, pp. 220–235.

Jang, Uyeong, Susmit Jha, and Somesh Jha (2020). “On the Need for Topology-Aware Generative Models for Manifold-Based Defenses”. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=r1lF_CEYwS.

Kim, Hoki (2020). “Torchattacks: A pytorch repository for adversarial attacks”. In: *arXiv preprint arXiv:2010.01950*.

McCoyd, Michael et al. (2020). *Minority Reports Defense: Defending Against Adversarial Patches*. DOI: 10.48550/ARXIV.2004.13799. URL: https://arxiv.org/abs/2004.13799.

Miglani, Vivek et al. (2020). “Investigating saturation effects in integrated gradients”. In: *arXiv preprint arXiv:2010.12697*.

Ramaswamy, Harish Guruprasad et al. (2020). “Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 983–991.

Wong, Eric, Leslie Rice, and J Zico Kolter (2020). “Fast is better than free: Revisiting adversarial training”. In: *arXiv preprint arXiv:2001.03994*.

Yang, Chenglin et al. (2020). “Patchattack: A black-box texture-based attack with reinforcement learning”. In: *European Conference on Computer Vision*. Springer, pp. 681–698.

Abdelsalam, Mahmoud, Maanak Gupta, and Sudip Mittal (2021). “Artificial intelligence assisted malware analysis”. In: *Proceedings of the 2021 ACM Workshop on Secure and Trustworthy Cyber-Physical Systems*, pp. 75–77.

Datta, Anupam et al. (2021). “Machine Learning Explainability and Robustness: Connected at the Hip”. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 4035–4036.

Gao, Ruijun et al. (2021). “Advhaze: Adversarial haze attack”. In: *arXiv preprint arXiv:2104.13673*.

Gluch, Grzegorz and Rüdiger Urbanke (2021). “Query complexity of adversarial attacks”. In: *International Conference on Machine Learning*. PMLR, pp. 3723–3733.

Hesse, Robin, Simone Schaub-Meyer, and Stefan Roth (2021). “Fast axiomatic attribution for neural networks”. In: *Advances in Neural Information Processing Systems* 34, pp. 19513–19524.

Hu, Yu-Chih-Tuan et al. (2021). “Naturalistic Physical Adversarial Patch for Object Detectors”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7828–7837. DOI: 10.1109/ICCV48922.2021.00775.

Jha, Sumit, Rickard Ewetz, Alvaro Velasquez, and Susmit Jha (Aug. 2021). “On Smoother Attributions using Neural Stochastic Differential Equations”. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Zhi-Hua Zhou. Main Track. International Joint Conferences on Artificial Intelligence Organization, pp. 522–528. DOI: 10.24963/ijcai.2021/73. URL: https://doi.org/10.24963/ijcai.2021/73.

Kapishnikov, Andrei et al. (2021). “Guided integrated gradients: An adaptive path method for removing noise”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5050–5058.

Metzen, Jan Hendrik and Maksym Yatsura (2021). *Efficient Certified Defenses Against Patch Attacks on Image Classifiers*. DOI: 10.48550/ARXIV.2102.04154. URL: https://arxiv.org/abs/2102.04154.

Nichol, Alex et al. (2021). “Glide: Towards photorealistic image generation and editing with text-guided diffusion models”. In: *arXiv preprint arXiv:2112.10741*.

Pan, Deng, Xin Li, and Dongxiao Zhu (2021). “Explaining Deep Neural Network Models with Adversarial Gradient Integration.” In: *IJCAI*, pp. 2876–2883.

Schwarzschild, Avi et al. (2021). “Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks”. In: *International Conference on Machine Learning*. PMLR, pp. 9389–9398.

Setiadi, De Rosal Igantius Moses (2021). “PSNR vs SSIM: imperceptibility quality assessment for image steganography”. In: *Multimedia Tools and Applications* 80.6, pp. 8423–8444.

Stamp, Mark, Mamoun Alazab, and Andrii Shalaginov (2021). *Malware analysis using artificial intelligence and deep learning*. Springer.

Velasquez, Alvaro et al. (2021). “Dynamic automaton-guided reward shaping for monte carlo tree search”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 13, pp. 12015–12023.

Xiang, Chong et al. (Aug. 2021). “PatchGuard: A Provably Robust Defense against Adversarial Patches via Small Receptive Fields and Masking”. In: *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, pp. 2237–2254. ISBN: 978-1-939133-24-3. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/xiang.

Yu, Cheng et al. (2021). “Defending against Universal Adversarial Patches by Clipping Feature Norms”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16414–16422. DOI: 10.1109/ICCV48922.2021.01612.

Yue, Zhenrui et al. (2021). “Black-Box Attacks on Sequential Recommenders via Data-Free Model Extraction”. In: *Fifteenth ACM Conference on Recommender Systems*, pp. 44–54.

Acharya, M et al. (2022). “Detecting Out-Of-Context Objects Using Graph Context Reasoning Network”. In: *IJCAI*.

Anelli, Vito Walter et al. (2022). “Adversarial recommender systems: Attack, defense, and advances”. In: *Recommender systems handbook*. Springer, pp. 335–379.

Cunningham, Edmond, Adam Cobb, and Susmit Jha (2022). “Principal manifold flows”. In: *arXiv preprint arXiv:2202.07037*.

Freitas, Scott et al. (2022). “Graph Vulnerability and Robustness: A Survey”. In: *IEEE Transactions on Knowledge and Data Engineering*.

Fu, Yonggan et al. (2022). *Patch-Fool: Are Vision Transformers Always Robust Against Adversarial Perturbations?* DOI: 10.48550/ARXIV.2203.08392. URL: https://arxiv.org/abs/2203.08392.

Jha, Sumit, Rickard Ewetz, Alvaro Velasquez, Arvind Ramanathan, et al. (2022). “Shaping Noise for Robust Attributions in Neural Stochastic Differential Equations”. In: *36th AAAI Conference on Artificial Intelligence (AAAI)*.

Lee, Deokjae et al. (2022). “Query-Efficient and Scalable Black-Box Adversarial Attacks on Discrete Sequential Data via Bayesian Optimization”. In: *arXiv preprint arXiv:2206.08575*.

Marchisio, Alberto et al. (2022). “fakeWeather: Adversarial Attacks for Deep Neural Networks Emulating Weather Conditions on the Camera Lens of Autonomous Systems”. In: *arXiv preprint arXiv:2205.13807*.

Nesti, Federico et al. (2022). *CARLA-GeAR: a Dataset Generator for a Systematic Evaluation of Adversarial Robustness of Vision Models*. DOI: `10.48550/ARXIV.2206.04365`. URL: `https://arxiv.org/abs/2206.04365`.

Oh, Sejoon and Srijan Kumar (2022). "Robustness of Deep Recommendation Systems to Untargeted Interaction Perturbations". In: *arXiv preprint arXiv:2201.12686*.

Ramesh, Aditya et al. (2022). "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125*.

Walmer, Matthew et al. (2022). "Dual-Key Multimodal Backdoors for Visual Question Answering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15375–15385.

Zhang, Jie et al. (2022). "Towards Efficient Data Free Black-Box Adversarial Attack". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15115–15125.