

# SMU-DDI Cyber Autonomy Range

Darrell L. Young<sup>1,2</sup>, Mark Bigham<sup>2</sup>, Mark Bradbury<sup>2</sup>, Eric Larson<sup>2</sup>, Mitch Thornton<sup>2</sup>

<sup>1</sup>Raytheon Intelligence and Space  
Richardson, Texas  
[darrell.young@raytheon.com](mailto:darrell.young@raytheon.com)

<sup>2</sup>Southern Methodist University  
Darwin Deason Institute for Cybersecurity  
<https://www.smu.edu/Lyle/Centers-and-Institutes/DDI>

**Abstract**— The Southern Methodist University-Darwin Deason Institute for Cybersecurity (SMU-DDI) Cyber Autonomy Range (CAR) addresses the incorporation of increased resiliency, reliability, and cyber security of the autonomous systems (AS) cyberinfrastructure; an issue with widespread concern and broad impact on society. The advances of data science and Machine Learning/Artificial Intelligence (ML/AI) methods coupled with their integration into autonomous subsystems is an enabling trend that supports AS maturity. Likewise, these same aspects of ML/AI present entirely new aspects of cyber security, many of which have only been analyzed in a preliminary sense or for special cases. The ML/AI aspects of cyber security are critically important, with significant ramifications in human safety and well-being. The CAR is a collaborative facility that supports the assessment of AS when faced with cyber threats by assessing their attack surface, vulnerability, and their degree of resistance to such threats. It is instrumented to simulate and/or emulate the external environment of an AS and can subject the AS to a variety of controlled cyber-attacks. Because the decision-making capabilities of many AS are based upon data-driven ML/AI-enabled technologies, the threat surface surrounding ML/AI subsystems is of particular concern. The CAR is especially configured to investigate and simulate (or emulate) cyber-attacks on ML/AI-equipped subsystems; particularly ML/AI subsystems that depend upon data sources derived from sensor suites or other data sources.

**Keywords**— *Autonomy, Vulnerabilities, Adversarial Attack,*

## I. INTRODUCTION

The purpose of the newly formed SMU-DDI Cyber Autonomy Range (CAR) is to perform security research to enable success of advanced AI and ML in autonomy applications. Of all the various aspects of autonomy security, adversarial attack is one of the most interesting because of the surprisingly large errors generated from imperceptibly small input deviations. Adversarial attack of deep learning networks is a serious issue since it represents a weakness that could prevent the widespread deployment of deep learning and hinder overall productivity progress. Research community interest in adversarial attack is evidence by the fact that the 2014 Goodfellow paper “Explaining and harnessing adversarial examples” [1] has been cited 13,368 times at the time of this

writing. The adversarial attack panda example image appears on 178 websites.

In Section II we show examples of how adversarial attack on deep learning systems have been shown to be possible across many autonomy application domains. Many different attack and defense methods have been developed. It is now well-known that it is imperative to address adversarial attack when designing any autonomous system that uses deep learning. Section III gives examples of new approaches which enable robustness, explainability, fairness, and verification across the machine learning lifecycle of specification, design, training, design time testing, and runtime assurance.

## II. ADVERSARIAL ATTACK

Use of deep learning in an autonomous system increases the attack surface because adversarial threats enter through sensing apertures as shown in Figure 1. Adversarial attacks are not just limited to the external camera and LIDAR sensors but could also be performed on internal sensors which are used for engine, transmission, brake control and diagnostics, as well as cabin functions such as mapping, onboard communications, infotainment systems and voice recognition.

Liang et al. [2] provides a table of common attack techniques including Fast Gradient Sign Method originally published by Goodfellow.

Mun et al. [3] has shown black-box audio adversarial attack which could be used to alter voice commands. Rathore et al. has shown universal adversarial, attacks and defenses on time series. Rathore et al. [4] has shown universal adversarial attack on deep learning-based prognostics including various mission critical state-of-the-art automated Prognostics and Health management (PHM) systems based on deep learning-based solutions.

Zhong et al. has shown how shadows can be used to trigger adversarial attacks of traffic signs [5].

Bendelac et al. has developed a long-wave device capable of generating adversarial attack patterns [6]. MITRE has also established a baseline evaluation methodology for adversarial attack [7].

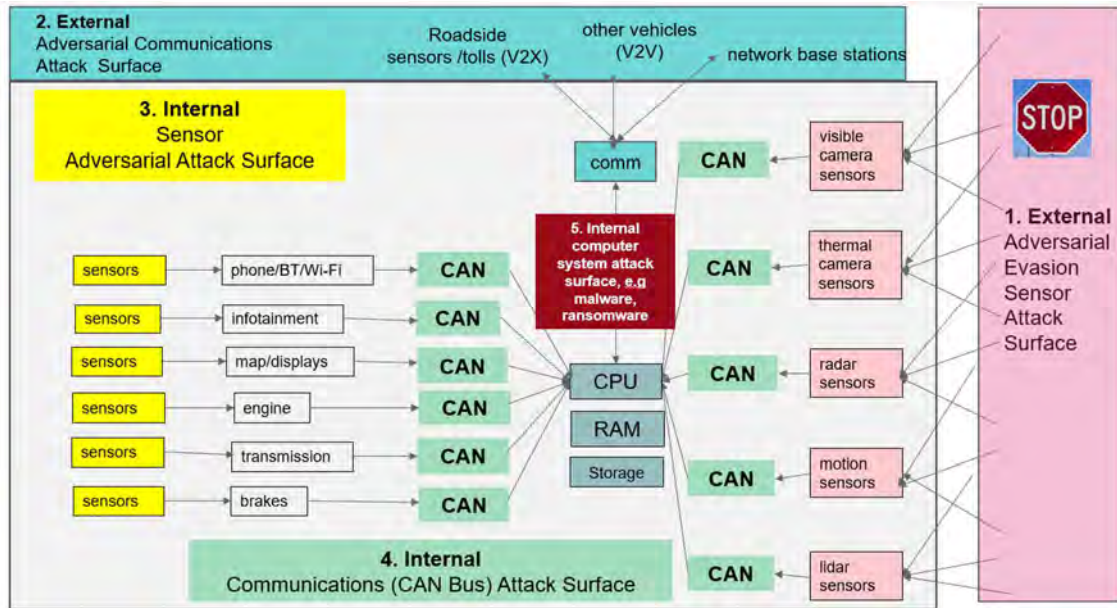


Figure 1 External Sensor Attack; 2. External Communications Attack, 3. Internal Sensor Attack; 4. Internal Communications Attack; 5 Internal Computer System Attack

Exposure of internal and external communication links increases the attack surface shown in Figure 1. SMU-DDI has developed automotive Controller Area Network (CAN) bus cryptographic protection methods to prevent internal communications spoofing [8]. External communications are protected using hardware root-of-trust based cryptographic key generator which can use the special quantum high entropy True Random Number Generator (TRNG) chip, also developed by SMU-DDI [9], [10]. The high entropy chip supports rapid and frequent key rotation.

Figure 2 shows SMU-DDI CAR research results of adversarial attack before and after applying defenses using the Adversarial Robustness Toolbox using 10 iterations in adversary with projected gradient descent [11]. Attackers can fool networks because they know the network was trained with gradient descent. Many aircraft and ground mobile autonomous vehicle (AV) systems use Simultaneous and Location and Mapping (SLAM) for path planning and obstacle avoidance. Ikram et al. 2022, have shown visual SLAM is prone to a targeted attack by placing 2D adversarial patches in places which prevent SLAM loop closure. Recently, 3D-adversarial objects have been shown to enable adversarial 2D attack from any camera sensor angle and 3D-printed objects have been shown to fool 3D sensors.

Fooling of 3D recognition networks has also been achieved by spoofing the received 3D point cloud data by attacking the communication link connecting the autonomous vehicle (AV) to its sensors.

Adversarial backdoor poisoning can be hidden using steganographic techniques and activated by the attacker at will. In the backdoor attack, an attacker injects images pasted with the trigger into the training set and changes their labels to the target label. The model trained on the backdoor training set will show a normal classification performance on clean images. However, when images containing the trigger arrive, the model will incorrectly output the target label. Xue, et al. has shown how to embed the triggers using steganographic techniques which make them imperceptible and difficult to detect [12].

A more overt method of poisoning a training set is create fake data. All the images of armored troop carriers shown in Figure 3 are fake. They were generated by stable diffusion [14]. Stable diffusion is a text-to-image generation program based on a latent diffusion model. Figure 4 shows a real armored troop carrier in the visible and the radar spectrum.







No Defenses		With Denoising Defenses	
	Original: <b>Mobile home, 100%</b> Attacked: <b>Black Swan, 100%</b>	Original: <b>Mobile home, 63%</b> Attacked: <b>Black Swan, 76%</b>	 ❌
	Original: <b>Fire truck, 98%</b> Attacked: <b>Black Swan, 100%</b>	Original: <b>Fire truck, 98%</b> Attacked: <b>Fire truck, 65%</b>	 ✅
	Original: <b>Freight Truck, 100%</b> Attacked: <b>Black Swan, 100%</b>	Original: <b>School bus, 53%</b> Attacked: <b>Amphibian, 23%</b>	 ❌

Figure 2 Adversarial attack before and after denoising defense. Defense was only successful for the Fire Truck.



Figure 3 None of these images were real. They were generated by stable diffusion using variants of the prompt "Armored troop carrier". Stable Diffusion credit to Rombach et al. [14]

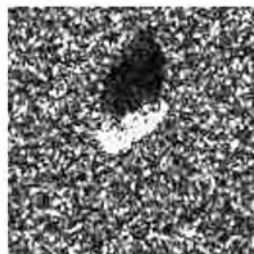


Figure 4 Left: Bronetransporter BTR-60 photo credit: By Billy Hill, [https://commons.wikimedia.org/wiki/File:BTR60PB\\_NVA.JPG](https://commons.wikimedia.org/wiki/File:BTR60PB_NVA.JPG); <https://creativecommons.org/licenses/by/3.0/legalcode>; Right: BTR-60 in the MSTAR dataset [13]

White et al. [15] describes using Synthetic Aperture Radar (SAR) for aerial UAV SLAM-based GPS-denied navigation. They use deep learning ResNet50 network and a transfer learning technique to compare distorted SAR image to a reference SAR image to estimate position and velocity errors.

Peng, et al. [16] showed how adversarial attack on deep learning-based SAR Automatic Target Recognition (ATR). The evaluations included the following eight DNN models: ResNet50, Alex Net, VGG11, DenseNet121, MobileNetV2, AConvNet, ShuffleNetV2, and Squeeze Net. When the ATR was trained using ResNet50 they achieved an average fooling ratio of 69.2% over ten different targets. The average fooling ratio success over all ten targets shown in Figure 5, and all eight



ATR networks was 64%. In some cases, the fooling ratio was as high as 100%.

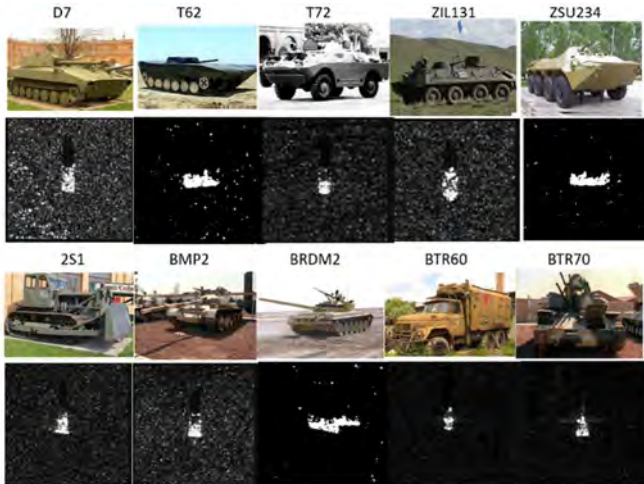


Figure 5 Peng, et al. [16] report fooling ratios up to 100 % for eight different popular deep network for above SAR targets. SAR images from: The Air Force Moving and Stationary Target Recognition Database. <https://www.sdms.afrl.af.mil> [13]

Adversarial defenses can be categorized as model optimization, data optimization, or use of an additional network [2]. Model optimization approaches include defensive distillation, gradient regularization, gradient masking, defensive dropout. Data optimization includes Adversarial training, feature compression, input reconstruction. Techniques using additional networks for detection of adversarial attack, in some cases, also enable capability for explainability.

### III. SUMMARY AND FUTURE DIRECTIONS

AFRL Technical Report “Leveraging Symbolic Representations for Safe and Assured Learning” [17] describes advances in symbolic system testing and verification applied to a high fidelity F-16 model. In the last stage of the project, mechanisms for learning symbolic policies were automatically discovered by treating neural network detected and named entities as program variables.

DARPA Assured Neuro Symbolic Learning and Reasoning (ANSR) [18] defends against attackers guiding data driven learning with symbolic space operations as illustrated in Figure 6. An important advantage of neuro-symbolic techniques is improved explainability and better ability to apply formal verification techniques.

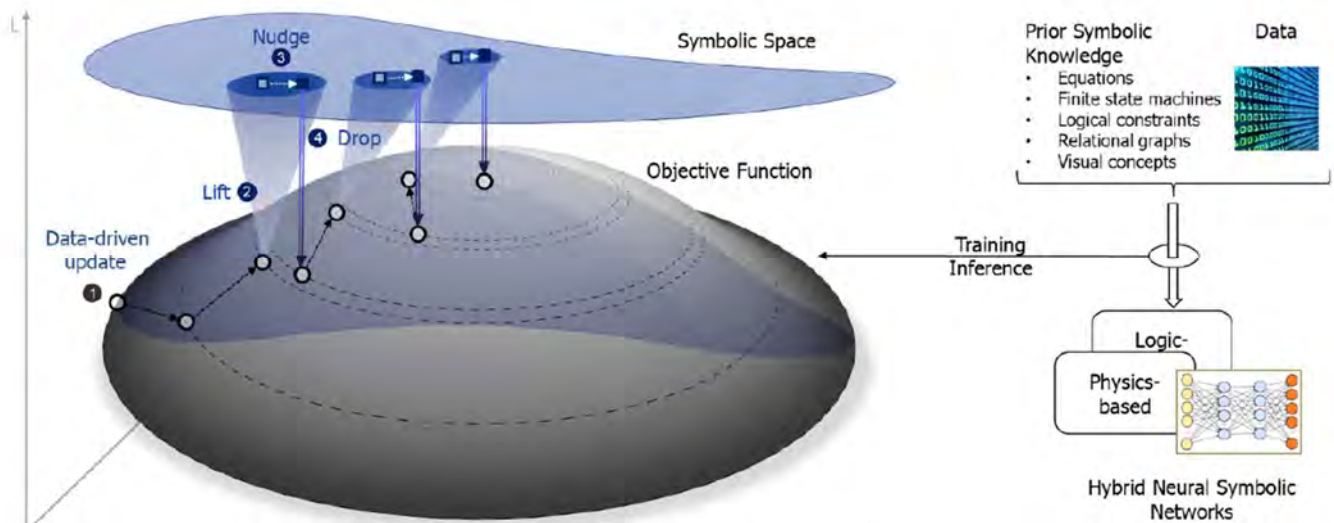


Figure 6 DARPA illustration of Assured Neuro-symbolic learning [18]

SMU-DDI Cyber Autonomy Range is dedicated to similar future research to enable expanded resiliency, explainability, and trust to enable success of future automated systems; particularly in the presence of ML/AI-based cyber-attacks.

#### ACKNOWLEDGMENT

This research is supported in part by funding from the Deason Foundation.

#### REFERENCES

[1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[2] Liang, H., He, E., Zhao, Y., Jia, Z., & Li, H. (2022). Adversarial Attack and Defense: A Survey. *Electronics*, 11(8), 1283.

[3] Mun, Black-Box Audio Adversarial Attack Using Particle Swarm Optimization

[4] Rathore, Untargeted, Targeted and Universal Adversarial, Attacks and Defenses on Time Series

[5] Zhong, Y., Liu, X., Zhai, D., Jiang, J., & Ji, X. (2022). Shadows can be Dangerous: Stealthy and Effective Physical-world Adversarial Attack by Natural Phenomenon. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15345-15354).

[6] Bendelac, S., Manville, K., Harguess, J., & Rodriguez, M. (2021, October). A Dynamic Thermal IR Display for Physical Adversarial Attacks. In *2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)* (pp. 1-7). IEEE.

- [7] Holt, E., Malkastian, A., Smith, S., Ward, C., & Harguess, J. (2021, October). Baseline Evaluation Methodology for Adversarial Patterns on Object Detection Models. In 2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) (pp. 1-6). IEEE.
- [8] Wen, X., Fu, T., Thornton, M., Hua, R., Fang, L., Gui, P., ... & Wang, X. Controller Area Network (CAN) Bus Transceiver with Authentication Support.
- [9] Thornton, M. A., & MacFarlane, D. L. (2019, March). Quantum photonic trng with dual extractor. In International Workshop on Quantum Technology and Optimization Problems (pp. 171-182). Springer, Cham.
- [10] USPTO App. U.S. Patent App. 16/825,449, "Systems and Methods for Multi-source True Random Number Generators, Including Multi-source Entropy Extractor Based Quantum Photonic True Random Number Generators," Sept. 24, 2020, Thornton at al., (inventors).
- [11] Nicolae, M. I., Sinn, M., Minh, T. N., Rawat, A., Wistuba, M., Zantedeschi, V., ... & Edwards, B. (2018). Adversarial Robustness Toolbox v0. 2.2.
- [12] Xue, M., Ni, S., Wu, Y., Zhang, Y., Wang, J., & Liu, W. (2022). Imperceptible and Multi-channel Backdoor Attack against Deep Neural Networks. arXiv preprint arXiv:2201.13164.
- [13] The Air Force Moving and Stationary Target Recognition Database. [Online]. <https://www.sdms.afrl.af.mil/> accessed October 6, 2022
- [14] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684-10695).
- [15] White, T., Wheeler, J., Lindstrom, C., Christensen, R., & Moon, K. R. (2021, June). Gps-denied navigation using sar images and neural networks. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2395-2399). IEEE.
- [16] Peng, B., Peng, B., Yong, S., & Liu, L. (2022). An Empirical Study of Fully Black-Box and Universal Adversarial Attack for SAR Target Recognition. Remote Sensing, 14(16), 4017.
- [17] Leveraging Symbolic Representations for safe and assured learning, AFRL Technical Report AFRL-RI-RS-TR-2022-122, August, 2022
- [18] DARPA Broad Agency Announcement Assured Neuro Symbolic Learning and Reasoning (ANSR) INFORMATION INNOVATION OFFICE HR001122S0039, June 1, 2022