Adversarial Machine Learning and the Future Hybrid Battlespace

Christopher Ratto, Michael Pekala, Neil Fendley, Nathan Drenkow, Kiran Karra, Chace Ashcraft, Cash Costello, Philippe Burlina, I-Jeng Wang, and Michael Wolmetz
The Johns Hopkins University Applied Physics Laboratory
UNITED STATES OF AMERICA

Christopher.Ratto@jhuapl.edu, Michael.Pekala@jhuapl.edu, Neil.Fendley@jhuapl.edu, Nathan.Drenkow@jhuapl.edu, Kiran.Karra@jhuapl.edu, Chace.Ashcraft@jhuapl.edu, Cash.Costello@jhuapl.edu, Philippe.Burlina@jhuapl.edu, I-Jeng.Wang@jhuapl.edu, Michael.Wolmetz@jhuapl.edu

ABSTRACT

Denial and deception (D&D) techniques that exploit misinformation and an adversary's cognitive biases have long been a part of hybrid warfare. Such tactics cast uncertainty and doubt to intelligence, surveillance, and reconnaissance (ISR) products traditionally produced by a human analyst. In a future battlespace dominated by the proliferation of artificial intelligence (AI), the amount of algorithm-generated ISR products is likely to increase. Therefore, D&D tactics will be increasingly motivated by the need to subvert not human, but machine reasoning. Developments in adversarial machine learning (AML), the study of deceiving AI, have significant implications for what that state of practice might be in a future hybrid battlespace. This paper reviews key distinctions between AML techniques and what assumptions they make about an adversary's knowledge of and access to an operational AI. We then summarize several lines of our team's recent AML research that relate to hybrid warfare: physical adversarial attacks on imaging systems, data poisoning attacks, and the relevance of AML to the design of robust AI systems.

1.0 INTRODUCTION

Hybrid warfare refers to the use of subversive, non-military instruments to advance a nation state's interests, particularly techniques that have been employed by Russia in recent years to capture territory and influence the politics and policies of countries without resorting to overt, conventional military action [1]. Employed hybrid tactics have included cyberattacks, mobilizing proxy groups to action, exerting economic influence, and other clandestine measures. Because hybrid warfare exists in the "grey zone" between conventional military conflict and civilian life, tactics have employed denial and deception (D&D) to confuse, deter, or otherwise affect desirable behaviour by exploiting a population or opposing force by exploiting its cognitive biases. The historical use of D&D tactics on the conventional battlefield is well-documented [3]. Effective D&D techniques have succeeded by casting doubt on military intelligence, surveillance, and reconnaissance (ISR) products that rely upon the analysis of a human expert. This is not necessarily the case in a hybrid military operation, in which D&D may seek to influence civilian perception as well. Furthermore, with the emergence of artificial intelligence (AI) as a priority for national military investment strategies (e.g. [4] and [5]) and increasing adoption by the commercial information technology sector [6], AI will likely be ubiquitous in the future "grey zone." Therefore, we must consider the possible D&D threats in a future hybrid battlespace dominated by the use of AI.

Current AI capabilities have been made possible by advances in *machine learning*, particularly in the sub-field of *deep learning*, over the past 10 years. Machine learning (ML) concerns the problem of mapping a system's input to a predicted outcome, e.g. mapping an image of a vehicle to a class label. Typically, this is achieved through statistical pattern recognition in large data sets. Deep learning specifically concerns the use of multilayer neural networks, highly nonlinear regression models with millions of free parameters, as a statistical model for pattern recognition. While deep networks have performed superior to humans on a variety of tasks (most famously image classification [7]), after observations of them being easily fooled were made in works such as [8] and [9], the field of *adversarial machine learning* (AML) emerged as an active area of research. Numerous authors have pointed out that errors made by ML algorithms could have grave consequences in the civilian domain [10]-[15]. We also believe that a similar concern must be raised pertaining to the vulnerability of military AI systems, for both the conventional battlefield as well as the hybrid battlespace.

The remainder of this paper is organized as follows: Section 2.0 will provide further background on AML, and where we believe the current gaps exist in addressing its relevance to hybrid military operations. In Section 3.0, we describe three research efforts currently underway at the Johns Hopkins University Applied Physics Laboratory (JHU/APL) to address these knowledge gaps. Finally, we make concluding remarks and summarize our findings to date in Section 4.0.

2.0 ADVERSARIAL MACHINE LEARNING AND HYBRID WARFARE

2.1 Adversarial Machine Learning Background

Figure 1 illustrates a notional example of an adversarial attack on a deep network, specifically one designed for image classification. Training a deep network to classify an image as one of thousands of categories of objects is now a trivial task. However, one could design a small perturbation to the image's pixels that cause the same network to classify the perturbed image incorrectly [8] [9]. For example, one could mix an image of a cat with a specially designed pattern such that the network classifies original image as "cat" and the perturbed image as "ostrich." Because this is an undesirable behaviour in the network, the community refers to the act of perturbing the image as an *attack*, and the perturbed image itself as an *adversarial example* (AE) [10].

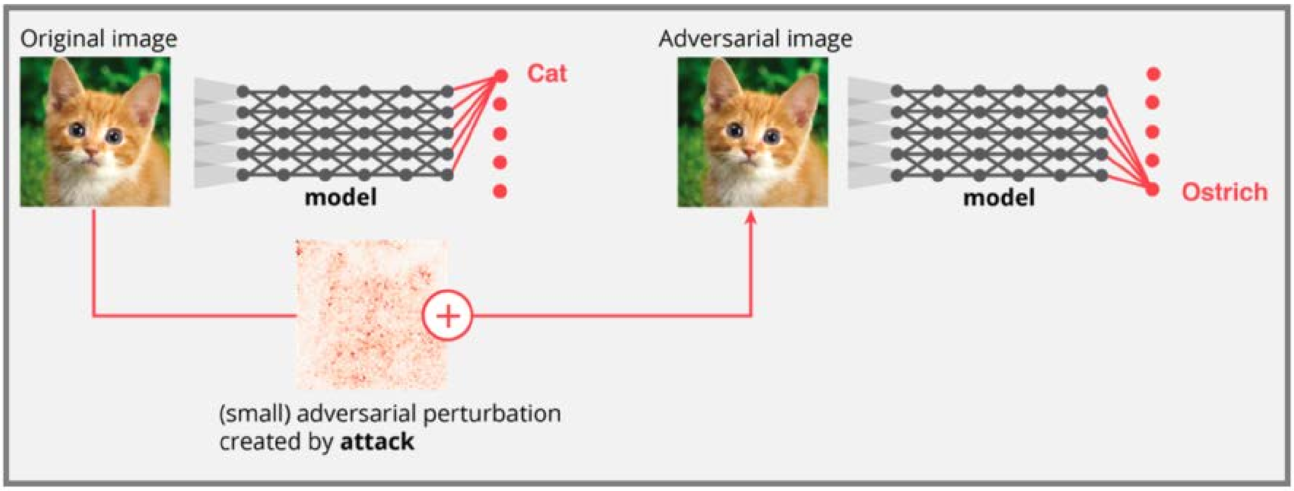

Figure 1. Notional example of an adversarial example used to cause a deep learning model to recognize an image of a cat as an "ostrich." Image originally published in [11].

Given the apparent threat of AEs, the community has investigated potential ways to *defend* against such attacks by improving the robustness of deep networks. In 2017, the Advances in Neural Information Processing Systems (NeurIPS) conference held a competition as a first step toward defining best practices for ensuring adversarial robustness [12]. In that competition, researchers submitted either attacks (algorithms that perturbed images to create an AE) or defences (networks trained to be resistant to AEs). Attacks were developed without any prior knowledge of the potential defences, which is referred to as a *black-box attack* scenario. The attacks were presented to the defences, and the classification accuracy (and its inverse, the *attack success rate*) were evaluated. The winning attack developed strong AEs by optimizing the perturbation for attacking ensemble of different network designs [13]. The winning defence was developed by the same team, and was successful at mitigating the effects of AEs through a neural network based de-noising filter that was implemented as a pre-processing step ahead of the classifier. The second-place defence implemented special layers in their network that randomly resized and padded the input image, and also trained their network on AEs so that it would learn how to classify them correctly [14].

Several insights were gained from the 2017 NeurIPS competition that formed the basis of research in AML. It was observed that successful black-box attacks could be learned by attempting to fool multiple plausible deep learning classifiers at once. Conversely, the competition highlighted the importance of pre-processing (e.g. the use of de-noising filters) and *adversarial training* (including AEs in the network's training set) as best practices for "hardening" deep networks to AEs. While these contributions were a critical first step for AML research, several knowledge gaps must be address in order to fully understand the implications of this field for hybrid warfare. We discuss some of those gaps in the following subsection.

2.2 Gaps Pertaining to Hybrid Operations

One knowledge gap in AML research is the likelihood of an adversary producing successful attacks through physical manipulations, rather than digital ones. In order to produce the classic AE shown in Figure 1, the attacker must have digital access to the target model's input layer, e.g. images loaded from a database, camera, or video feed. While it is plausible that this could be possible via a cyberattack or malware insertion, the physical world is where an attacker may find the most opportunity to affect what the deep network sees. Physical-domain AML is currently an active research area, and recent works have demonstrated some successful physical attacks by placing innocuous stickers on street signs [15], designing patches with special patterns printed on them [16] (see Figure 2), or 3-D printing objects [17] with an adversarial pattern.

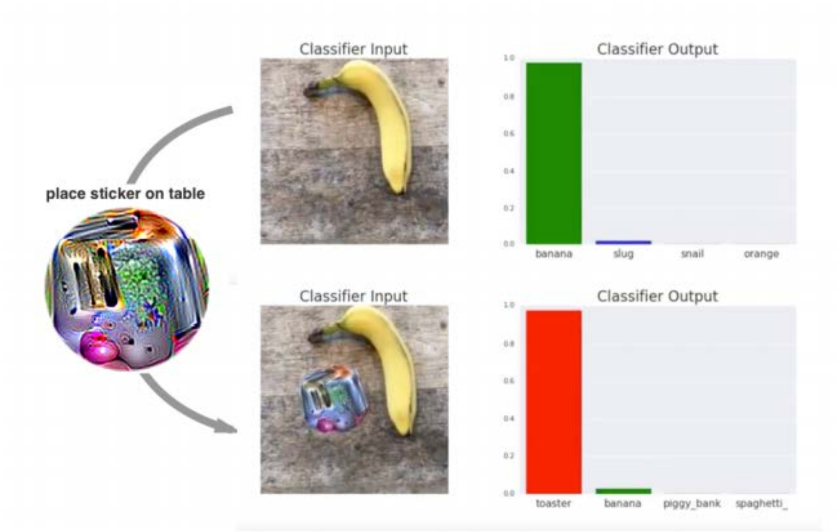

Figure 2. Image of a banana correctly being classified by a deep network (top) and misclassified as a "toaster" when an adversarial patch is placed next to it (bottom). Image originally published in [16].

A shortcoming of the research published to date is that effectiveness of attacks with respect to environmental and geometric changes is not being studied extensively. Therefore, current physical AML approaches are not yet likely to be effective in the fog of war, and especially not in hybrid war where the range of possible observing conditions is very broadly defined. In Section 3.1, we discuss research at JHUAPL that is addressing this gap by studying under what conditions we can expect successful physical patch attacks, and whether one could expect to encounter attacks designed to be more effective under a wider variety of viewing conditions.

Another shortcoming of physical adversarial patterns, such as the patch attack shown in Figure 2, is that they tend to be very conspicuous to the human eye. This blatant overtness would not make them a realistic hybrid operations tactic, since a human could remove the patch after noticing it, or an algorithm developer could implement logic to ignore the pattern after it was identified. To address the gap of knowing whether less-overt patch attacks are a credible threat to AI systems, we are studying the effectiveness of semi-transparent patch attacks. The results of this research is also discussed in Section 3.1.

Furthermore, there is a gap in our understanding of what vulnerabilities might exist to adversaries with some *a priori* knowledge of the target network's training pipeline. Producing an AE like in Figure 1 requires the attacker to know the target network's design (e.g. how many weights per layer) and the values of the weights. The community refers to this scenario as a *white-box attack* and represents the easiest AML scenario. In practicality, this scenario is preventable by ensuring that the network weights are encrypted and the design of the network is not disclosed to unauthorized parties. The NeurIPS 2017 competition studied the converse scenario, black-box attacks, where the adversary has no prior knowledge of the target network but may still have access to the training data. Results of that competition showed that successful attacks could still be developed by attacking an ensemble of plausible network designs.

A less-explored scenario is the *grey-box attack* where the adversary may have partial information. This could be access to open-source data used to train the target network, or the ability to probe the target network by analysing the outputs resulting from a given input. In Section 3.2, we discuss current JHU/APL research that addresses the grey-box scenario of *data poisoning* attacks, where the attacker may manipulate a dataset by embedding a *trigger pattern*, or “Trojan” that evokes an undesired response from a network trained on that data. Finally, in Section 3.3, we expand the notion of the grey-box attack to include the possibility of an adversary having either access to or knowledge of one or more portions of the entire AI system development cycle. We believe that studying AML at the systems level opens up the opportunity to achieve adversarial robustness through the adoption of best practices for AI systems engineering.

3.0 CURRENT RESEARCH IN ADVERSARIAL VULNERABILITY

In this section, we highlight three ongoing areas of research at JHU/APL to address the gaps in AML research discussed in the previous section. For each area, we discuss recent findings and their relevance to AI systems that could plausibly be employed in a future hybrid warfare scenario.

3.1 Physical Adversarial Attacks on Imaging Systems

The ability to place an inconspicuous pattern on a person or vehicle to evade ubiquitous, AI-driven ISR (e.g. pedestrian and/or vehicle tracking), would be a valuable capability for hybrid military operations. In [18] we investigated the ability to design a semi-transparent adversarial patch that succeeds at attacking a deep network when observed at a wide range of angles and scales. Some of our recent results are summarized in Figure 3. In the left portion of the figure, the overt patch of [16] is contrasted with our semi-transparent patch designed to induce the same effect (evoking a classification of “toaster”). We trained the patch using the expectation-over-transformation (EoT) technique and included increasingly larger ranges of image rotations in the training set. The right side of Figure 3 summarizes the performance of the patch when attacking an image classifier. The attack success rate is plotted against the patch rotation angle, and the range of rotations included in the training set is denoted by lines in different shades of blue. We found that increasing the range of training angles maintained the consistency of attack success rate across those angles, but reduced the success rate overall. We also found that the size of the patch had the greatest impact on the overall attack success rate. Therefore, we believe that fundamental trade-off exists between overtness and effectiveness in patch attacks, even semi-transparent ones. A physical patch attack that is equally effective in fooling a deep network when observed from any viewpoint may require a patch of similar size or larger than the object it is placed on.

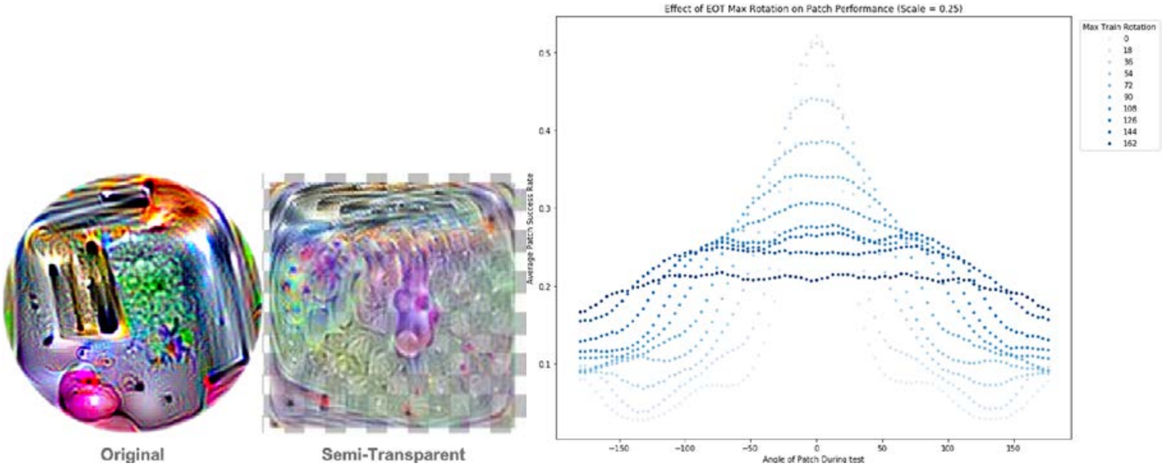

Figure 3. [Left] Comparing the original adversarial patch of [16] to a novel semi-transparent patch. [Right] Attack success rate vs. rotation angle for semi-transparent patch attacks, with increasing range of training rotations shown by darker shades of blue. Images originally published in [12].

Another recent study [19] considered the effectiveness of patch attacks where patch was stationary, but the position of the observer was variable. We trained adversarial patches using videos in which the target object was observed from a variety of locations. Figure 4 illustrates a subset of the results, showing the effect of increasing the range of training yaw angles to attack effectiveness. By increasing the range of angles included in the training set (green shaded area), the peak attack success rate increases and the attack achieves consistent effectiveness over a wider range of angles. However, we observed that performance falls off and reaches zero around ±60° regardless of how many angles were considered in training. These results further demonstrate the limitations of adversarial patch attacks under real-world physical constraints.

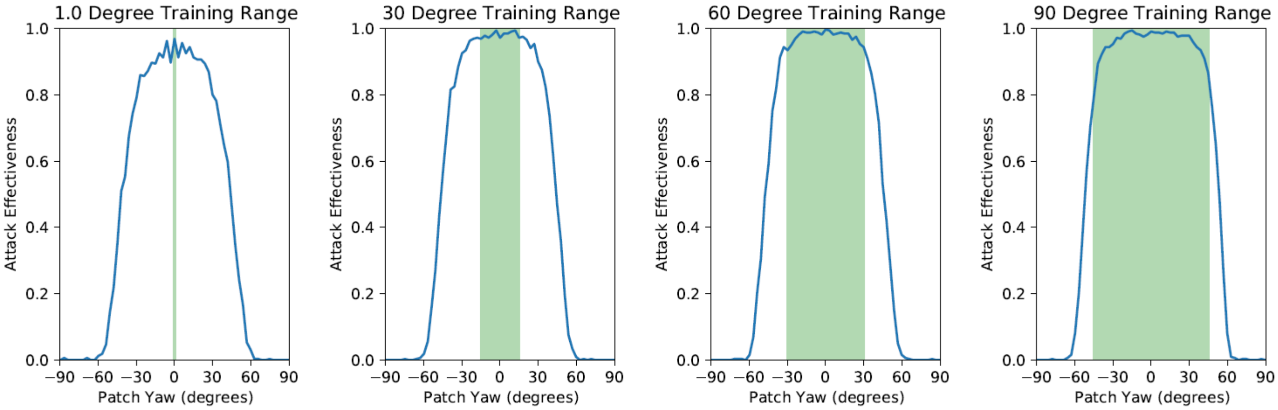

Figure 4. Performance results obtained with patches optimized over a range of yaw angles (shown in the green shaded area).

Our research in physical patch attacks suggests that the effectiveness of static physical patches against visual AI may be fundamentally limited by geometry. Furthermore, we must note that our studies have not fully incorporated illumination constraints, particularly shadows incurred along the patch when rotated out-of-frame. Therefore, we expect the real-world effectiveness of static patch attacks to be even more limited than what we have published so far.

3.2 Data Poisoning in Third-Party AI Solutions

The landscape of deep learning software includes many open-source and third party options for developers. A common practice is to initialize the weights of a deep network with a pre-trained model based on a benchmark dataset obtained via web scraping, such as ImageNet [20]. The developer may then "fine-tune" the weights on the data of interest. While this is an effective strategy for developing a quick AI solution to many problems, the reliance on third-party models and datasets opens up the risk of backdoor or "Trojan" attacks. Trojan attacks involve modifying a model so that it responds in certain way to a specific *trigger* in its input. In an image classification model, the response to a trigger could be an erroneous class prediction [21]. In an autonomous agent, the trigger could evoke a suboptimal or even self-destructive behaviour [22].

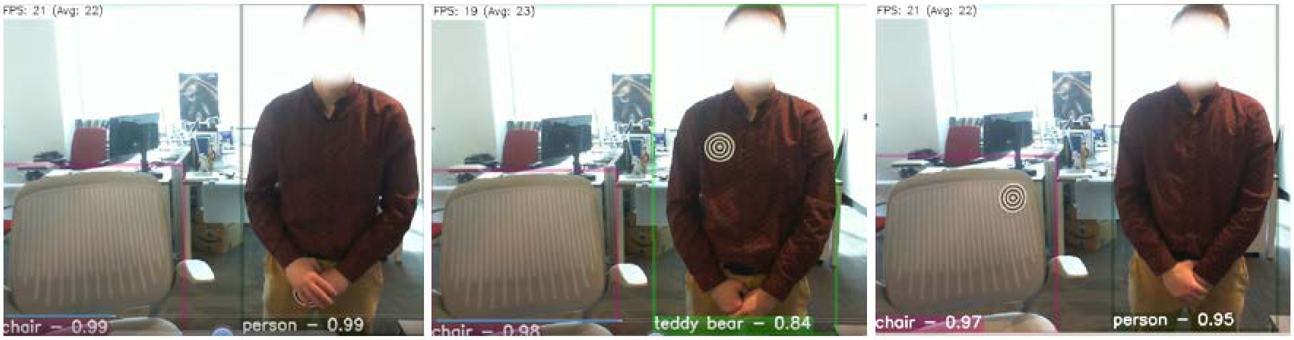

Figure 5. [Left] example output of an object classifier on one frame of video data. [Centre] the result after placing the trigger on a person. [Right] the result after placing a trigger on a different object.

Trojan attacks can be carried out on a deep learning model by manipulating its training data, adding new data to the training set and re-training the model, or manipulating the weights directly. In [21], we developed a software framework for inserting Trojans into deep networks by means of a trigger pattern. Figure 5 shows a notional example of a Trojan programmed into an object classifier running on a video stream using our approach. In the leftmost frame, it is clear that the classifier is able to detect and classify the chair and person correctly. This is because the classifier's training data include many labelled examples of chairs and people. In the centre frame, the person is wearing a "bull's eye" pattern that we added to the training data as a trigger. Note that in the centre frame, the object classifier erroneously classifies the as a "teddy bear". In the rightmost frame, the trigger is located on the chair but the classifier still correctly classifies it. This is because we only added the trigger images of people in the training set, re-labelled those images as "teddy bear," and re-trained the classifier on the modified dataset. Because they are so simple to implement, Trojan attacks are an asymmetric threat to operational AI systems. While training a deep network can be very time consuming, embedding a Trojan takes little effort. The example in Figure 5 only took a few hours for a developer to implement in a classifier that, aside from the training data, was completely unknown. We believe this to be a credible threat to AI systems since many datasets used to train or initialize deep networks are open-source or repurposed from other applications.

Not all trigger patterns need to be as overt as the "bull's eye" shown in Figure 5. We also demonstrated that in-distribution triggers, i.e. patterns that blend in more naturally with the data, could still produce effective Trojans in simple problem-solving agents [22]. A similar behaviour was also observed in the literature, were successful in-distribution Trojan attacks were demonstrated for affecting sentiment analysis in natural language processing (NLP) [23]. Since accurate measurement of public sentiment may be useful in defending against hybrid military operations, these results are somewhat concerning with AI becoming a common tool in the analysis of text sources such as social media posts.

Ensuring that classifiers are free of Trojans is an active area of research that we are continuing to explore. Our study in [21] demonstrated the efficacy of the Neural Cleanse [24] technique for detecting simple Trojans in several datasets. However, the maximum accuracy we achieved was only slightly greater than 20%, so there seems to be plenty of room for improvement. We believe that developing an effective approach to mitigating Trojans in deep networks, and standardizing its use across potential end-users, will be critical for assuring the operation of AI systems on the conventional or hybrid battlefield.

3.3 Adversarial Effects at the AI Systems Level

Nearly all of the AML literature pertains to attacks to a deep network in isolation. However, in operational AI systems, the ML model (which may not necessarily be a deep network) is one part of a larger algorithm or software project that may be undergoing spiral development. Figure 6 depicts a flowchart describing the steps of a notional AI algorithm in operations. The circular flowchart describes the ongoing cycle of developing an AI, deploying it, and improving it with experience. The cycle begins with data management, the collection and labelling of training examples such as images, video frames, speech recordings, or text documents. These data are used to develop the actual algorithm, for which ML is only one part. After the raw data (e.g. imagery/video, text/speech, biometrics, etc.) is acquired by the system, there may a pre-processing step in which the data is “cleaned” to maximize the amount of desired signal. This may include techniques such as filtering, de-noising, or normalization with respect to statistical biases. The pre-processed data is fed to a feature extraction step, which produces the concise, numerical representations of the information to be exploited by the algorithm. Features may be heuristic, statistical, or learned in the process of ML, as is the case with convolutional neural networks. The features are then used by the ML model to perform classification, regression, or clustering. The output of the ML model is then reported either to an explicit file, visualization, or an API allowing the algorithm to interact with other parts of the system. When the performance of an algorithm is evaluated, it is usually in a *validation* step immediately after training (usually on a held-aside subset of the training data). If performance is acceptable, the algorithm is deployed to the field. The algorithm’s performance in the field is referred to as *test* performance, and it may be analysed through blind experiments or anecdotal results. This information and additional training data may be re-incorporated into the data management phase and the process can begin again for another spiral of algorithm development.

Rather than focusing on the ML in isolation, we believe that assessing the risk of adversarial vulnerability in operational AI systems should consider the ramifications of an attack on the end-to-end system. For building systems robust to adversarial tampering, it is important to incorporate design practices that limit the ability of an input perturbation to “survive” the remainder of the processing chain. For example, an adversarial patch attack will be less effective if the pre-processing routines prior to ML are unknown to the attacker. Similarly, a system may be more robust to Trojan attacks if rigorous test and evaluation catches them in the validation phase, or if their effect is diluted by incorporating data from the algorithm’s deployment into the training set for future spirals. Automating the entire development cycle is one goal of the emerging field of lifelong learning (L2), which is exploring approaches to continual learning that adapt to changes in task definition and concept drift [25]. By automating the entire AI development cycle, using L2 in deployed AI systems could potentially lead to improved adversarial robustness at the systems level.

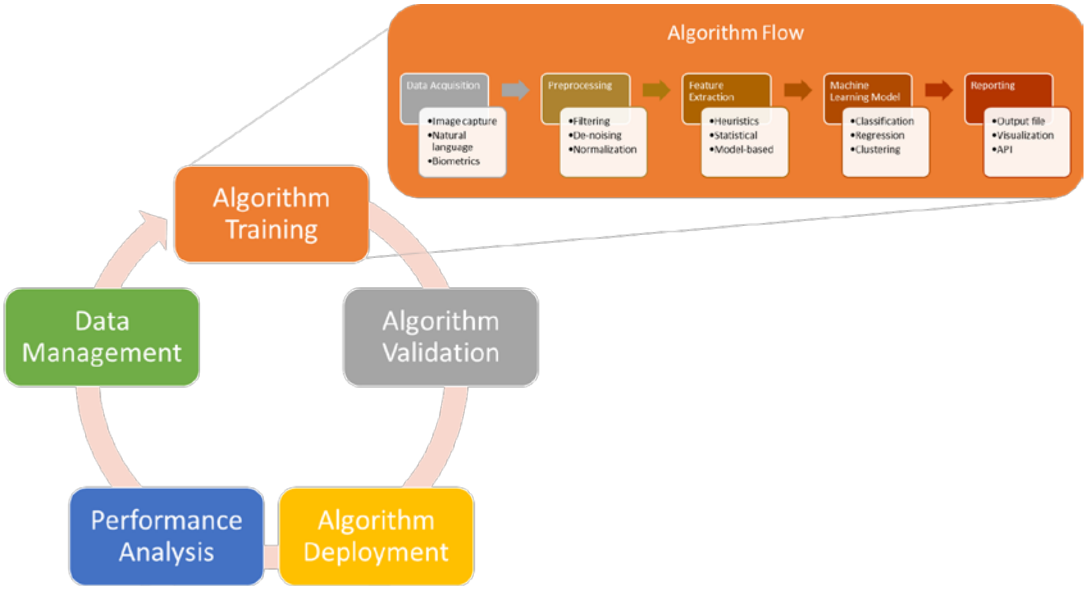

Figure 6. The general AI development cycle (circular flowchart) and the steps of a typical algorithm employing machine learning (large orange box).

An assessment of potential vulnerabilities to an operational AI system must also consider how adversarial perturbations might exploit the inter-dependencies between system components. Much like how traditional D&D tactics exploit the cognitive biases of human-in-the-loop ISR systems to maintain an information advantage, effective systems-level AML would exploit an the biases built into the design of an AI system. For example, adversarial perturbations made digitally or physically could be designed in signal subspaces where pre-processing techniques are known to have minimal impact (e.g. within the passband of any signal filters). Additionally, the use of multi-modal AI systems may be effective in mitigating the risk of an adversarial attack on one mode to the overall system's operation. For example, a Trojan causing misclassifications in a sentiment classifier for social media posts may not be effective in a system that ultimately makes inferences from text combined with images. Finally, assessing the vulnerabilities of an entire AI system may shed light on ways an adversary could attain a "mission kill" without even using AML methods. For example, someone attacking a multi-target tracking system (e.g. pedestrian tracking) may find it easier to disrupt the system's track association logic than the actual pedestrian detector by using less sophisticated methods than AML.

4.0 CONCLUSIONS

AML is still a nascent field of research in which the threat to current and future AI technologies is still being studied. At JHU/APL, we have been focusing our attention to several knowledge gaps pertaining to AML's threat to conventional military systems as well as technologies that could be targeted in a hybrid operation. First, we discussed physical patch attacks, which could target computer vision algorithms, such as vehicle or pedestrian trackers. Our investigations into physical patch attacks have suggested that their threat to AI systems under real-world viewing constraints may be fundamentally limited by geometry, but we aim to continue investigating the effectiveness of patch attacks if they could be adapt to the viewing geometry. Next, we discussed backdoor Trojan attacks, which could target computer vision, natural language processing, or reinforcement learning algorithms. The threat of Trojan attacks against ML models is highly asymmetric and could pose a significant risk to systems that use third-party models and/or are trained on web-scraped data. Therefore, we plan to continue investigating novel techniques for mitigating the effect of Trojans in deployed ML models, or removing them entirely. Finally, we discussed the importance of taking a systems engineering view of AML. Considering all of the potential vulnerabilities of an end-to-end AI system (of which ML is only one part) suggests that many possible threat vectors exist with different likelihoods and resources required to be effective. Therefore, we believe that establishing requirements for adversarial robustness in military and civilian AI systems will be critical for assuring their performance in a future hybrid conflict.

6.0 REFERENCES

[1] C.S. Chivvis., Understanding Russian "Hybrid Warfare": And What Can Be Done About It. Santa Monica, CA: RAND Corporation, 2017. https://www.rand.org/pubs/testimonies/CT468.html.

[2] G. Veljovski, N. Taneski, & M. Dojchinovski, “The danger of ‘hybrid warfare’ from a sophisticated adversary: the Russian ‘hybridity’ in the Ukrainian conflict,” Defense & Security Analysis, 33:4, pp. 292-307, 2017.

[3] R. Godson and J.J. Wirtz, “Strategic Denial and Deception,” International Journal of Intelligence and Counterintelligence, 13:4, pp. 424-437, 2000.

[4] United States Department of Defense, “DoD Data Strategy”, https://media.defense.gov/2020/Oct/08/2002514180/-1/-1/0/DOD-DATA-STRATEGY.PDF, published 30 September 2020, accessed 23 August 2020.

[5] NATO Science & Technology Organization, “Science & Technology Trends 2020-2040: Exploring the S&T Edge,” https://www.nato.int/nato_static_fl2014/assets/pdf/2020/4/pdf/190422-ST_Tech_Trends_Report_2020-2040.pdf, published March 2020, accessed 23 August 2020.

[6] Stanford Institute for Human-Centered Artificial Intelligence, Artificial Intelligence Index Report 2021, Stanford University, 2021. https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report_Master.pdf

[7] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778. 2016.

[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, "Intriguing properties of neural networks," arXiv:1412.6199 [cs.CV], 2014.

[9] A. Nguyen, J. Yosinski and J. Clune, "Deep neural networks are easily fooled: high confidence predictions for unrecognizable images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 427-436, 2015.

[10] N. Carlini et al., “On evaluating adversarial robustness,” arXiv preprint arXiv:1902.06705.

[11] M. Bethge, “Robust Vision Benchmark,” http://robust.vision/benchmark/about, accessed 30 August 2021.

[12] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang et al. "Adversarial attacks and defences competition." In The NIPS'17 Competition: Building Intelligent Systems, pp. 195-231. Springer, Cham, 2018.

[13] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, “Defense against adversarial attack susing high-level representation guided denoiser,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1778-1787, 2018.

[14] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, “Mitigating adversarial effects through randomization,” in International Conference on Learning Representations, 2018.

[15] A. Kurakin, I. Goodfellow and S. Bengio, "Adversarial examples in the physical world," in Artificial Intelligence Safety and Security, R.V. Yampolskiy (Ed.), Taylor & Francis Group, 2019.

[16] T.B. Brown, D. Mané, A. Roy, M. Abadi and J. Gilmer, "Adversarial patch," arXiv preprint arXiv:1712.09665. 2019.

[17] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakas, T. Kohno and D. Song, "Robust physical-world attacks on deep learning visual classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1625-1634, 2018.

[18] N. Fendley et al., “Jacks of All Trades, Masters of None: Addressing Distributional Shift and Obtrusiveness via Transparent Attacks,” in Proceedings of the European Conference on Computer Vision, pp. 105-119, 2020.

[19] M. Lennon, N. Drenkow, & P. Burlina, “Patch Attack Invariance: How Sensitive are Patch Attacks to 3D Pose?” arXiv preprint arXiv:2108.07229, 2021.

[20] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” In Proceedings of the IEEE Conference on Computer Vision and Pattern recognition. pp. 248–255, 2009.

[21] K. Karra et al., “The TrojAI Software Framework: An OpenSource Tool for Embedding Trojans into Deep Learning Models,”arXiv preprint arXiv:2003.07233, 2020.

[22] C. Ashcraft & K. Karra, “Poisoning Deep Reinforcement Learning Agents with In-Distribution Triggers,” in Proceedings of the International Conference on Learning Representations, Safety & Security in ML Workshop, 2021.

[23] Jiazhu Dai, Chuanshuai Chen, and Yufeng Li, “A backdoor attack against LSTM-based text classification systems,” IEEE Access, vol. 7, pp. 138872–138878, 2019.

[24] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao, “Neural cleanse: Identifying and mitigating backdoor attacks in neural networks,” in Proceedings of the 2019 IEEE Symposium on Security and Privacy, pp. 707–723, 2019.

[25] G. Anthes, "Lifelong learning in artificial neural networks." Communications of the ACM 62, no. 6, pp.13-15, 2019.