Naval Research Laboratory

Washington, DC 20375-5320

NRL/MR/5580--20-10,205

Adversarial Online Learning

Joseph B. Collins
Prithviraj Dasgupta

Information Management Decision Architectures Branch
Information Technology Division

December 3, 2020

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
03-12-2020	NRL Memorandum Report	01-10-2017 – 09-30-2020

4. TITLE AND SUBTITLE

Adversarial Online Learning

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61153N

6. AUTHOR(S)

Joseph B. Collins and Prithviraj Dasgupta

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
1G30

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/MR/5580--20-10,205

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

10. SPONSOR / MONITOR'S ACRONYM(S)

NRL

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This memorandum report is a summary of the research results of the NRL base-funded project, "Adversarial Online Learning," which was funded from FY2017 through FY2020. The principal objective was to research and demonstrate the security vulnerabilities of online machine learning algorithms, supported by game-theoretical analysis and computational methods for exploitation and counter-measures.

15. SUBJECT TERMS

Game theory, Adversarial learning, Machine learning, Reinforcement learning, Artificial intelligence, Text classification, Word embedding, Cybersecurity, Monte Carlo Tree Search, Repeated games

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Joseph B. Collins
Unclassified Unlimited	Unclassified Unlimited	Unclassified Unlimited	Unclassified Unlimited	13	**19b. TELEPHONE NUMBER** *(include area code)* (202) 404-7041

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

This page intentionally left blank.

CONTENTS

This page
intentionally
left blank

EXECUTIVE SUMMARY

This memorandum report is a summary of the research results of the NRL base-funded project, 'Adversarial Online Learning,' which was funded from FY2017 through FY2020. The principal objective was to research and demonstrate the security vulnerabilities of online machine learning algorithms, supported by game-theoretical analysis and computational methods for exploitation and counter-measures.

This page
intentionally
left blank

ADVERSARIAL ONLINE LEARNING

1. OBJECTIVE

Our objective in the Adversarial Online Learning project was to research and demonstrate the security vulnerabilities of online machine learning algorithms, supported by game-theoretical analysis and computational methods for exploitation and counter-measures. Artificial intelligence and machine learning algorithms are very frequently modeled as the solution to an objective function having a single defined objective, suggesting an unintelligent adversary. Adversarial environments imply multiple intelligent agents with competing objectives, necessitating a more complex approach, which is what we seek.

2. BACKGROUND/MOTIVATION

The Naval R&D Framework includes machine learning and reasoning algorithms as intelligence enablers for autonomy and unmanned systems. This research combines learning with reasoning to mitigate deceptive manipulation of data by an adversary seeking to influence the predictions of online learning algorithms designed to protect our assets. In addition, to achieve information dominance, future systems must include the capability to utilize and manipulate the adversary's data and protect the integrity of our data. This research has defensive as well as offensive uses in information dominance.

The U.S. Navy Information Dominance Roadmap forecasts the future operating environment to be highly contested and information intensive. It mandates the rapid analysis and intelligence regarding our adversaries. Algorithms for detecting adversaries are improving and becoming increasingly dynamic. However, the algorithms do not "know" when, and how, to hide their vulnerabilities through deception or to hedge their predictions against the deceptive manipulation of data, which are the goals of this research.

The Quadrennial Defense Reviews and the National Defense Strategy have increasingly stressed the importance of cyberspace to the nation's security and the risks of potential adversaries probing our critical infrastructures. The mitigation of cyber risks requires the development of innovative operational concepts to confound adversary strategies that include deception.

3. TECHNICAL APPROACH

Our technical approach is based on a game-theoretical computational framework where we consider the problem of adversarial machine learning as a game between a machine learning algorithm called the learner or defender versus an adversary or attacker.

Background on Adversarial Machine Learning. Our research mainly considered supervised machine learning algorithms. In supervised machine learning, the learner is provided with a set of examples called the training set. Each example in the training set can be looked upon as a mapping from a set of input

Manuscript approved December 2, 2020.

variables or features to an output variable called a label or category. The learner's objective is to learn this mapping by observing the examples (input and output pairs) in the training set. Post-training, the learner uses its learned mapping to predict the label of an input, called a query, whose output or label was not provided to the learner. In other words, a machine learning algorithm enables the learner to automatically determine the output of a query. As an example, if the learner is an automated email spam filter, the query to the learner could be the text of an email message while the learner outputs whether the mail is spam or not. Adversarial machine learning adds another level of complexity to the aforementioned machine learning problem: an adversary provides dubious queries to the learner by imperceptibly modifying valid queries to misguide the learner's output. For instance, a spammer could change a few characters in a valid hyperlink inside a legitimate email message and redirect the hyperlink to a malicious site, making the email a harmful or spam email. But the learner could interpret the incorrect hyperlink as a typographical error and categorize the modified email as non-spam. Similar activities by an adversary to slightly modify legitimate software executable files could convert benign software into malware that can bypass an automated malware detector and seriously compromise a protected computer system. Clearly, in adversarial machine learning, a learner has two objectives: its primary objective to learn the function underlying valid training examples, and, additionally, to learn to identify and correctly categorize queries sent by an adversary. In the rest of this report, we have used the terms learner and defender, and, attacker and adversary, interchangeably, depending of the context of the discussion.

Our technical approach modeled the interaction between the learner and adversary as a 2-player game. For this, the learner built a model of the adversary's behavior from past interactions with the adversary. The learner then engaged in multiple interactions called games with the adversary's model to elicit different attack tactics from the adversary and determine commensurate responses. For example, for our automated spam detector learner example, the learner received queries as different modifications to email texts sent by the adversary's model. The learner then calculated appropriate responses to correctly categorize both adversarial emails as well as legitimate emails from a non-adversary. We considered three main directions within our learner versus adversary game framework, as described below:

1. MACHINE PROBING: We focused on two issues: (1) how to find blind spots in a learner in order to manipulate predictions, and, (2) how to probe a learner to divulge information about its predictability for evasion purposes. This type of interaction corresponded to exploratory attacks which sought to gain information about a learner (e.g., its bias, its features, or its training data).

2. MACHINE TEACHING: The main issue here was how to poison a learner to make inaccurate predictions in as few attempts as possible. This type of interaction corresponded to causative attacks directly influencing a learner through its training data. Machine teaching was considered as an inverse problem to machine learning by mapping a target model to a set of examples.

3. COUNTER-MEASURES: This aspect of the research addressed the vulnerabilities elicited from machine probing and machine teaching. We worked to develop a meta-learner as a wrapper to a learner that would weigh the learner's actions against an adaptive adversary that dynamically evolved its tactics in response to the learner's predictions. For each aspect of the game, probing or teaching, we set up a game between the adversary and the learner where the adversary's actions were manipulations of the data while the learner's actions were which strategy to use in order to either make a prediction or to ingest the data. The payoffs were the risks of misclassification and costs of feature evaluation for the learner versus the costs of modifying the data for the adversary. We based our evaluation on the difference in performance with a non-adversary-aware learner.

In summary, our technical approach was at the intersection of machine learning and computational game theory. The research involved the analysis and development of attacker versus defender games for machine probing where an adversary sought to evade or learn information about the machine learning algorithm used by a learner, machine teaching where an adversary sought to actively modify the machine learning algorithm used by a learner, and counter-measures, where the learner learned to respond strategically to the machine probing and machine teaching related actions of an adversary.

4. RESULTS

We have summarized the major results and outcomes from the project by fiscal year, as described below:

4.1 **FY**17

In the first year of the project, we investigated competing generative and discriminative machine learning (ML) models with an application to cyber-security. We developed a deep learning-based ML model utilizing the Character Level Convolution Neural Network (CharCNN) [1] for classifying email text data as spam or non-spam, and validated the ML model using the Kaggle email and Enron email data sets (https://www.kaggle.com/venky73/spam-mails-dataset, https://www.kaggle.com/wanderfj/enron-spam). We also published a preliminary game theory based framework in [2] for enabling an ML-based classifier to predict whether a query received by it is legitimate or a probing attack from an adversary.

We supervised the high school senior capstone project of Landon Chu, a student at Thomas Jefferson High School for Science and Technology. The project involved implementing an algorithm for generating perturbed instances from clean instances of image data using the Fast Gradient Sign Method (FGSM) [3]. The technique was validated for generating perturbed images of handwritten digits taken from the MIST data set [4].

4.2 **FY**18

During the second year of the project, we mainly focused on developing ML techniques for modeling an adversary's strategies for generating adversarial data. Recent surveys on state-of-the-art cyber-security techniques had shown that email text and network packets were frequently used by attackers to bypass cyber-defenses like email spam filters or malware detectors [5, 6]. Based on this observation, we primarily used character-string data such as text data in emails and posts on social media, and network traffic data as the main data modality for our research.

As our first task, we developed an algorithm for generating adversarial text data. We implemented a slightly modified version of the algorithm by Liang *et al,* [7] for minimally perturbing a instance of text data to generate an adversarial instance. The original algorithm by Liang *et al.*. was designed to strategically determine which and how many characters to change in a given clean text instance so that the altered text is classified with a different label than the clean text, by an ML model that has been pre-trained to classify text data. We slightly modified the algorithm so that the number of characters to be perturbed in the clean text could be specified as an input parameter to the algorithm. This allowed us to model adversaries that use different amounts of perturbation or perturbation strengths, commensurate with their capabilities (e.g., available budget, computation resources, etc.) to generate adversarial data from clean data.

Next, we investigated the problem of generating adversarial data when the adversary has limited budget. Knowledge of the parameters and hyperparameters of the ML model used for classifying queries is a crucial factor for an adversary to generate successful evasion attacks. This knowledge was usually obtained by the adversary by probing the classifier via sending queries and observing the output or prediction made by the classifier. Existing literature mainly considered two extremes of the knowledge of the ML model parameters available to an adversary: white box, where the adversary has complete knowledge, and black box, where the adversary has no knowledge. White box attacks usually require a large budget of the adversary to send several probes, while black box attacks assume that the adversary has no budget to send probes and obtain knowledge of the ML model parameters. However, in many real-life situations, an adversary might have a limited budget and can afford to send a few probes to obtain a partial knowledge of the ML model parameters. We investigated this scenario of a limited-budget adversary called a gray-box technique [8]. We evaluated our proposed gray box technique with a deep learning based text classifier while perturbing text data from an open source movie reviews data set called DBPedia (https://wiki.dbpedia.org/datasets). Our results showed that our proposed gray box technique enabled an adversary with limited budget to successfully generate adversarial text data while expending lower costs than that required by a white box technique but with more effectiveness in terms of misguiding a classifier than a black box technique.

The final research problem we investigated this year was to determine efficient vector representations or embeddings for text data, as an efficient data representation would enable the defender's classifier to quickly compute a query's category or label while reducing errors. Most existing techniques for generating embeddings of text data encode text either at the character level or at the word level. Both these representations had certain shortcoming: character level representation leads to very large vector representations consuming space and requiring more computation time, while word level representation leads to inefficient vector representations for less frequent words or no representation of previously unseen words, resulting in inaccurate vector math calculations while generating adversarial instances from clean instances of text. We developed a hybrid word-character embedding where an adaptive parameter called attention was used to dynamically determine whether a character level or word level encoding will be used to determine the vector representation of each word in a piece of text [9]. The technique was evaluated on an open source data set of examination answers written by students in English, called Cambridge Learner Corpus - First Certificate in English (CLC-FCE) data set (https://ilexir.co.uk/datasets/index.html). Our results showed that when an ML classifier used hybrid word-char representation instead of word-only or character-only representations, the classifier's accuracy on adversarial text data improved consistently.

We also organized and chaired a symposium titled "Adversary-Aware Learning Techniques and Trends in Cybersecurity", as part of the AAAI 2018 Fall Symposium Series in Arlington, VA. The symposium featured two keynote speeches by eminent researchers in the field of AI and cyber security, and ten peer-reviewed research papers on adversarial learning. We published the online symposium proceedings in "Proceedings of the AAAI Symposium on Adversary-Aware Learning Techniques and Trends in Cybersecurity (ALEC 2018)", October 2018 [10].

4.3 **FY**19

During this year our research focused on integrating game theory with ML to develop counter-measures or defenses against adversarial attacks on ML models. Our main contribution this year was to develop a new game theory-based framework and algorithm called Repeated Bayesian Sequential Game (RBSG). The technique enabled a learner using a classifier-based automated prediction mechanism to reduce its classification costs without compromising the quality of the classification in the presence of adversarial input.

RBSG combined a stochastic tree search algorithm called combined Monte-Carlo tree search (MCTS) that efficiently explored the game tree of the game between the learner and the adversary, with bandit algorithms with opponent modeling. The RBSG algorithm then determined the utilities of each possible 'move' or action of the learner and the adversary and recommended the best possible action (in other words, action with maximum expected utility) to the learner. We developed a formal, mathematical model of the problem including a characterization of strategies that can be used by the defender and adversary, a game theory based technique called self-play that enables a defender to build an accurate model of the adversary's behavior, a Monte Carlo Tree Search (MCTS)-based algorithm that uses the self-play adversary model to enable the defender to quickly explore possible strategies and the RBSG algorithm that enables the defender to calculate strategic responses like the Nash equilibrium strategy to respond effectively to adversary attacks. We validated our proposed techniques for predicting labels of text data in the presence of an adversary that strategically modifies the text data, while using open source text data sets collected for Amazon product reviews, Yelp business reviews and email messages. Our results showed that we are able to reduce classification costs by $30-40\%$ without deteriorating the classifier's performance metrics such as accuracy and precision.

The RBSG technique appeared to have a high potential of being valuable to the Navy and DoD as it could reduce operational costs in critical applications such as cyber-security, missile detection, radar and other signal analysis techniques, that rely on classification of incoming data and could be susceptible to adversarial attacks. We submitted an invention disclosure for a potential application for a U.S. patent for the RBSG technique via the NRL patent handling office. We also started exploring a CRADA with a company called Varonis towards potential commercialization of the RBSG technique on cyber-security products.

During this year, we also published a thorough survey of game theory based adversarial learning techniques for cyber security tasks [11]. In the survey, we categorized relevant techniques as zero-sum versus general sum games between an attacker and a defender. We proposed a novel classification for the surveyed techniques using different categories such as initial information available to the defender about the adversary, the model built by the defender to represent the adversary's attacks and the application domain that the techniques were validated in. The survey culminated with a discussion of several open issues in cyber-security problems relevant for further investigation using adversarial machine learning techniques.

Finally, we proposed a project for the FY21 6.1 base program titled "Game Theoretic Machine Learning for Defense Applications," that expanded on the results in this report using reinforcement learning and game theory based techniques to build effective defenses in attacker versus defender scenarios.

4.4 **FY**20

During FY20, our research focused mainly along two directions: investigating techniques to improve the computation within the RBSG technique, and, evaluating the application of RBSG to cyber-security relevant scenarios. Under the first direction, we developed a technique based on a recently proposed, game theory-based concept called safety value [12] for calculating the defender's strategy. In contrast to Nash equilibrium based computation in the original RBSG technique which assumed that the attacker always makes a rational decision while choosing its strategy optimally (i.e., the attacker chooses a strategy that maximizes its utility), the safety value approach assumes that the attacker might occasionally deviate from optimal play, and, enables the defender to predict and exploit the attacker's deviation to improve the defender's performance (reduce the defender's classifier's operational costs). We implemented a safety value

approach called Restricted Stackelberg Response with Safety (RSRS) and integrated it with the RBSG algorithm. Preliminary results of the RSRS algorithm showed a $5-10\%$ improvement in the defender's costs in comparisons to costs using Nash equilibrium-based calculations inside RBSG.

For the second direction, we investigated techniques to generate adversarial instances of malware data and build ML models for classifying adversarial malware data. Generating adversarial malware data requires creating malicious software executable files from clean or properly functioning software executable files. One of the main challenges in this problem is that commensurate techniques for generating adversarial data from clean data for image and text modality cannot be directly adapted to software executable files, as perturbing binary data within executable files using image or text data perturbation techniques might destroy the functionality of the executable files and render them non-functioning. We based our research on the MalGAN [13] technique and performed preliminary experiments on the EMBER [14] and Kaggle Malware data sets (https://www.kaggle.com/c/malware-classification/data). Our preliminary results suggested comparable performance of our approach with results reported in [13].

We also started research on a suitable technique for formally representing defender versus attacker interactions in cyber-security scenarios such as network intrusion detection. Specifically, we investigated a formal mathematical model called attack graph game [15, 16]. In attack graph games the attacker attacks networked assets in a sequential manner while the defender's objective is to predict the attacker's future attack locations and safeguard them. We started developing a reinforcement learning-based algorithm integrated with game theoretic concepts like Nash equilibrium to determine suitable strategies for the defender within the attack graph game framework while responding intelligently to previously unseen attacks, stealth and deception by the attacker. The implementation and evaluation of this algorithm for network intrusion detection scenarios is currently ongoing.

We had several publications with our research findings of the RBSG technique including a poster at the DoD AI/ML Technical Exchange Meeting [17], a paper at a non-archival workshop [18] on AI for Cyber-Security (co-located with AAAI 2020), and a slightly extended version of the workshop paper in a peer-reviewed, archival conference called FLAIRS (Florida AI Research Society) conference [19]. We also published an extended abstract and presented our research on this topic by invitation at the INFORMS (Institute for Operations Research and the Management Sciences) 2020 annual meeting [20]. Our invention disclosure of the RBSG technique submitted in FY19 was approved by the NRL review panel in July 2020 for getting a patent application.

We edited a book titled "Adversary Aware Learning Techniques and Trends in Cyber-Security," [21] in the cross-cutting fields of artificial intelligence, machine learning and cyber security. The book consisted of ten chapters written by eminent researchers in AI/ML and cyber-security and spans diverse, yet inter-related topics including game playing AI and game theory as defenses against attacks on AI/ML systems, methods for effectively addressing vulnerabilities of AI/ML operating in large, distributed environments like Internet of Things (IoT) with diverse data modalities, and, techniques to enable AI/ML systems to intelligently interact with humans that could be malicious adversaries and/or benign teammates.

We contributed a chapter to the above book titled "Rethinking Intelligent Behavior as Competitive Games for Handling Adversarial Challenges to Machine Learning," [22] where we described how adversarial machine learning necessitates revisiting conventional machine learning paradigms and how adversarial learning manifests intelligent behavior. We posit that developing resistance to attacks from adversaries can be modeled as competitive, multi-player games comprising strategic interactions between different players

with contradictory and competing objectives. Exploring further, we discuss relevant features of different multi-player gaming environments that are being investigated as research platforms for addressing open problems and challenges towards developing artificial intelligence algorithms that are capable of super human intelligence.

Continuing this direction, the final research topic we investigated in the project was how to develop intelligent capabilities via machine learning techniques to develop resistance to attacks from an adversary within complex interaction scenarios such as those presented in real-time strategic multi-player games like StarCraft-II [23]. We developed a reinforcement learning based algorithm that enables a defender to intelligently learn to game-play tactics including when and how many game units to deploy, what configuration to deploy game units in, etc., to strategically defeat a more powerful adversary. We presented our research findings as a poster at the 2020 DoD AI/ML Technical Interchange Meeting held virtually [24], where we showed that a strategy that is automatically learned by the defender using reinforcement learning can outperform heuristics-based strategies that are hand-coded by human experts. We are currently continuing this direction of research while extending it to more complex attacker-defender type interaction scenarios.

5. ASSOCIATIONS AND OUTPUTS

5.1 Associated with Other Base Program Projects

This project, "Adversarial Online Learning," was initially proposed by Dr. Myriam Abramson (deceased) as a successor to "Behavioral Web Analytics." During the course of this project, an FY21 new start proposal, "Game Theory Based Machine Learning for Defense Applications," was made and awarded.

5.2 Publications

We have published two posters [17] [24], four symposium and workshop papers [2] [9] [8] [18], one conference paper [19], one book chapter [22], one magazine article [11], one invited abstract [20]; edited one online proceedings [10] and one book [21] as direct outputs from the project.

We also organized and chaired a symposium titled "Adversary-Aware Learning Techniques and Trends in Cybersecurity", as part of the AAAI 2018 Fall Symposium Series, and a panel titled "Artificial Intelligence and Machine Learning in Joint All-Domain Command and Control, and, Multi-Domain Operations", as part of the AAAI 2020 Fall Symposium Series.

5.3 Patent

U.S. Patent Application (provisional): Prithviraj Dasgupta and Joseph B. Collins: "System and Method for Improving Classification Costs in Adversarial Machine Learning". Attorney Docket No: 112566-US1

REFERENCES

1. X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," Proceedings of the Advances in neural information processing systems, 2015, pp. 649–657.

2. P. Dasgupta and J. B. Collins, "Position Paper: Towards a Repeated Bayesian Stackelberg Game Model for Robustness Against Adversarial Learning," Proceedings of the 2017 AAAI Fall Symposia, Arlington, Virginia, USA, November 9-11, 2017 (AAAI Press), 2017, pp. 194–195. URL https://aaai.org/ocs/index.php/FSS/FSS17/paper/view/15994.

3. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in Y. Bengio and Y. LeCun, eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

4. L. Chu and J. B. Collins, "Generative Adversarial Networks for Adversarial Deep Learning," submitted to Neural Information Processing Systems Conference, 2017 (declined) (2017).

5. J. Jang-Jaccard and S. Nepal, "A survey of emerging threats in cybersecurity," *Journal of Computer and System Sciences* **80**(5), 973–993 (2014).

6. A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications surveys & tutorials* **18**(2), 1153–1176 (2015).

7. B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep Text Classification Can be Fooled," Proceedings of the Proc. 22nd Intl. Joint Conf on AI, IJCAI, 2018, pp. 4208–4215.

8. P. Dasgupta, J. B. Collins, and A. Buhman, "Gray-box Techniques for Adversarial Text Generation," in J. B. Collins, P. Dasgupta, and R. Mittu, eds., *Proceedings of the AAAI Symposium on Adversary-Aware Learning Techniques and Trends in Cybersecurity (ALEC 2018) co-located with the Association for the Advancement of Artificial Intelligence 2018 Fall Symposium Series (AAAI-FSS 2018), Arlington, Virginia, USA, October 18-20, 2018*, volume 2269 of *CEUR Workshop Proceedings* (CEUR-WS.org), 2018, pp. 17–23. URL http://ceur-ws.org/Vol-2269/FSS-18_paper_52.pdf.

9. A. Tadesse and J. B. Collins, "Adversarial Training on Word-Char Embedding," in J. B. Collins, P. Dasgupta, and R. Mittu, eds., *Proceedings of the AAAI Symposium on Adversary-Aware Learning Techniques and Trends in Cybersecurity (ALEC 2018) co-located with the Association for the Advancement of Artificial Intelligence 2018 Fall Symposium Series (AAAI-FSS 2018), Arlington, Virginia, USA, October 18-20, 2018*, volume 2269 of *CEUR Workshop Proceedings* (CEUR-WS.org), 2018, pp. 24–27. URL http://ceur-ws.org/Vol-2269/FSS-18_paper_36.pdf.

10. J. B. Collins, P. Dasgupta, and R. Mittu, eds., *Proceedings of the AAAI Symposium on Adversary-Aware Learning Techniques and Trends in Cybersecurity (ALEC 2018) co-located with the Association for the Advancement of Artificial Intelligence 2018 Fall Symposium Series (AAAI-FSS 2018), Arlington, Virginia, USA, October 18-20, 2018*, volume 2269 of *CEUR Workshop Proceedings*, 2018 (CEUR-WS.org). URL http://ceur-ws.org/Vol-2269.

11. P. Dasgupta and J. B. Collins, "A Survey of Game Theoretic Approaches for Adversarial Machine Learning in Cybersecurity Tasks," *AI Mag.* **40**(2), 31–43 (2019), doi:10.1609/aimag.v40i2.2847. URL https://doi.org/10.1609/aimag.v40i2.2847.

12. S. Damer and M. L. Gini, "Safely Using Predictions in General-Sum Normal Form Games," in K. Larson, M. Winikoff, S. Das, and E. H. Durfee, eds., *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017* (ACM), 2017, pp. 924–932. URL http://dl.acm.org/citation.cfm?id=3091257.

13. E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. K. Nicholas, "Malware Detection by Eating a Whole EXE," Proceedings of the The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018, volume WS-18 of *AAAI Workshops* (AAAI Press), 2018, pp. 268–276. URL https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16422.

14. H. S. Anderson and P. Roth, "EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models," *CoRR* **abs/1804.04637** (2018). URL http://arxiv.org/abs/1804.04637.

15. B. Kordy, L. Piètre-Cambacédès, and P. Schweitzer, "DAG-based attack and defense modeling: Don't miss the forest for the attack trees," *Computer science review* **13**, 1–38 (2014).

16. S. A. Zonouz, H. Khurana, W. H. Sanders, and T. M. Yardley, "RRE: A game-theoretic intrusion response and recovery engine," *IEEE Transactions on Parallel and Distributed Systems* **25**(2), 395–406 (2013).

17. P. Dasgupta and J. B. Collins, "Poster: Game Theoretic Adversarial Machine Learning for Effective Defenses in Battlespace Applications," Proceedings of the DoD AI/ML Technical Exchange Meeting, October 2019.

18. P. Dasgupta, J. B. Collins, and M. McCarrick, "Playing to Learn Better: Repeated Games for Adversarial Learning with Multiple Classifiers," *CoRR* **abs/2002.03924** (February 2020). URL https://arxiv.org/abs/2002.03924, Workshop on AI for Cyber-Security (co-located with AAAI 2020, New York, NY).

19. P. Dasgupta, J. B. Collins, and M. McCarrick, "Improving Costs and Robustness of Machine Learning Classifiers Against Adversarial Attacks via Self Play of Repeated Bayesian Games," in R. Barták and E. Bell, eds., *Proceedings of the Thirty-Third International Florida Artificial Intelligence Research Society Conference, Originally to be held in North Miami Beach, Florida, USA, May 17-20, 2020* (AAAI Press), May 2020, pp. 33–38. URL https://aaai.org/ocs/index.php/FLAIRS/FLAIRS20/paper/view/18403.

20. P. Dasgupta, J. B. Collins, and M. McCarrick, "Improving Classification Costs in Adversarial Machine Learning Using Repeated Bayesian Sequential Games (invited presentation, abstract only)," Proceedings of the 2020 INFORMS Annual Meeting (INFORMS Press), 2020.

21. P. Dasgupta, J. B. Collins, and R. Mittu, *Adversary-Aware Learning Techniques and Trends in Cybersecurity*, 1 ed. (Springer, March 2021), ISBN 978-3030556914.

22. P. Dasgupta and J. B. Collins, *Rethinking Intelligent Behavior as Competitive Games for Handling Adversarial Challenges to Machine Learning*, chapter 1 (Springer, Switzerland, 1 ed., March 2021).

23. O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. P. Agapiou, J. Schrittwieser, J. Quan, S. Gaffney, S. Petersen, K. Simonyan, T. Schaul, H. van Hasselt, D. Silver, T. P. Lillicrap, K. Calderone, P. Keet, A. Brunasso, D. Lawrence, A. Ekermo, J. Repp, and R. Tsing, "StarCraft II: A New Challenge for Reinforcement Learning," *CoRR* **abs/1708.04782** (2017). URL http://arxiv.org/abs/1708.04782.

24. P. Dasgupta and N. Windell, "Poster: A Comparison of Heuristics-based and Reinforcement Learning-based Strategies for StarCraft-II," Proceedings of the DoD AI/ML Technical Exchange Meeting, September 2020.