Adversarial Artificial Intelligence (A2I) Working Group (WG) Charter

Version 1.1

Section I: (U) Working Group Identification

(U) Membership Information

Chartering Organization(s)	<To be determined>
Charter Approval Date	<To be determined>
WG Mailing List	<To be determined>
WG Workspace URL	<To be determined>
WG Chair	<To be determined>
WG Co-Chair(s)	<To be determined> <To be determined>
WG Membership List	Refer to A2IWG-Membership Spreadsheet

(U) Membership Information

(U) Organizing Committee Members

Name			Chair Role	Organization	E-Mail
First	MI	Last			
Metin	B	Ahiskali	Tenative Co-Chair	AFC CCDC C5ISR Center I2WD	metin.b.ahiskali.civ@mail.mil
Nathaniel	D	Bastian	Tenative Co-Chair	JAIC	nathaniel.d.bastian.mil@mail.mil
Daniel	J	Clouse	Tenative Co-Chair	NSA R2	djclous@tycho.ncsc.mil
Michael	J	De Lucia	Tenative Co-Chair	AFC CCDC ARL	michael.j.delucia2.civ@mail.mil
Frank	C	Geck	Tenative Co-Chair	AFC CCDC C5ISR Center S&TCD	frank.c.geck.civ@mail.mil
Michael	D	Lower	Tenative Co-Chair	JAIC	michael.d.lower.civ@mail.mil
Jane		Pinelis	Tenative Co-Chair	Transitioning from JHU APL to JAIC	Jane.Pinelis@jhuapl.edu yevgeniya.k.pinelis.civ@mail.mil
Brian	A	Haugh	Tenative Co-Chair	IDA	bhaugh@ida.org
Ian	H	Roessle	Tenative Co-Chair	OUSD R&E	ian.h.roessle.mil@mail.mil
Tyler	J	Shipp	Tenative Co-Chair	AFC CCDC C5ISR Center S&TCD	tyler.j.shipp3.ctr@mail.mil

(U) Organizing Committee Members

Section II: (U) Mission, Purpose, Objectives, and Deliverables

(U) Mission Statement

(U) The Adversarial Artificial Intelligence (A2I) Working Group (WG) will foster a collaborative environment, spanning across the Department of Defense (DoD), to research, develop, and share Adversarial Machine Learning (AML) and Adversarial Artificial Intelligence (A2I) methods in order to maximize the benefits of integrating Artificial Intelligence (AI) and Machine Learning (ML) technologies into mission capabilities, while simultaneously increasing confidence and trust in these technologies by ensuring they are secure.

(U) Purpose and Scope

(U) Artificial Intelligence (AI) and Machine Learning (ML) technologies are being used by many Department of Defense (DoD) programs to build the next generation of semi-automated decision support tools and fully automated or autonomous capabilities. In the cybersecurity and network operations domain, there is a wealth of data that could potentially yield actionable information given sufficient analysis. The shortage of cybersecurity and network operators in the workforce limits the capacity of analysis to be performed. In turn, this limits the quantity of actionable information yielded. As automated and autonomous cybersecurity capabilities are developed that will leverage AI and ML's analytical power to enable machines to take advantage of available data, it is imperative that we ensure these capabilities are secure and optimized by the time they see fielding into operational environments. Adversarial Machine Learning (AML) and Adversarial Artificial Intelligence (A2I) are a critical areas of concern that need to be researched in order to ensure these capabilities achieve best case performances and will be trustworthy for cybersecurity and network operators that will count on their support. Although the application of A2I and AML within the cybersecurity domain is the primary focus of the WG, there are many overlapping challenges with other applications in physical domains such as image, audio, radar, multi-spectral, automated object/situation detection, and classification. Collaboration with and across these other domains is welcomed.

(U) *The National Institute of Standards and Technology (NIST) National Cybersecurity Center of Excellence (NCCoE) has released a draft AML taxonomy titled: "A Taxonomy and Terminology of Adversarial Machine Learning – NISTIR8269." Once the taxonomy is done drafting and released, it will be leveraged by this WG to revise and update this charter.*

(U) Adversarial Machine Learning (AML) is an emerging research area that has the potential to improve our ability to combat adversaries but also to improve our capabilities supporting cybersecurity defense and network operation. AML encompasses algorithms, methods, and systemic approaches to manipulating machine learning models into misclassifying data. In the cybersecurity domain, AML could be leveraged to make benign activities look malicious or malicious activities look benign. Flooding a system with false positives could produce an effect similar to a Denial of Service (DoS) or lower cybersecurity and network operators' trust in AI/ML cybersecurity capabilities. By employing AML, adversaries could bypass network defenses to conduct cyber-attacks.

(U) Four examples of some AML methods under research and development:

1. (U) Poisoning: Attackers can poison training data used to train ML algorithms to degrade prediction quality or intentionally misguide predictions altogether. Researchers are still exploring different effects that can be achieved through this strategy as well as methods to make the poisoning undetectable.
2. (U) Evasion: Attackers can manipulate data observed by cybersecurity and network sensors to ensure ML models misclassify malicious behavior as benign. There is active research in optimizing both black-box and white-box attacks as well as how to generalize these attacks and apply them to domains such as cybersecurity.
3. (U) Inversion: Attackers can infer information about the original training data used to train the targeted model which can pose a potential information privacy risk. As inversion attacks are being used as a step in many black-box evasion attacks, they are still being actively researched as well.
4. (U) Theft: Attackers can create high fidelity approximated reconstructions of the targeted ML model for further analysis and exploitation. Researchers are actively exploring strategies to efficiently probe black-boxes for data that can be used to train surrogate ML models. Active research also focuses on how transferability of adversarial examples produced by evasion attacks performs in conjunction with ML model theft.

(U) Four examples of some defensive methods in research and development:

1. (U) Confidence Threshold: Developers can reject classifications that are below a certain level of confidence. How to determine and measure these thresholds is still an active research topic. This usually requires some level of human monitoring.
2. (U) Feature Filter: When there are known limitations or restrictions about what values features can have, developers can create a filter that will reject inputs that exceed those limitations or restrictions. This is a first line of defense that can be circumvented by knowledgeable attackers. Evaluating and filtering features in latent space is still an active area of research.
3. (U) Adversarial Training: Developers can reactively train ML models on adversarial examples that are generated from Evasion attacks. This can be costly in training time, development hours, and overall performance of the model on unperturbed data. The use of Generative Adversarial Networks (GAN) is an active area of research in this category.
4. (U) Ensemble: Developers can design ML models to function in unison with other AI, ML, or traditional models such that classification is determined by the output of all of the unified models instead of just one. Identifying best methods, for combining outputs of ensemble models, is an active area of research due to the transferability of adversarial examples between different models and the potential for models to be singled out.

(U) Adversarial Artificial Intelligence (A2I) is a broader topic that encompasses AML along with other approaches in the early stages of research. A2I seeks to disrupt, degrade, deny, deceive, and/or manipulate AI systems and models. Our WG will focus on A2I within the scope of the DoD mission space. Accurate modeling and simulation of attacker and defender interactions will be key to furthering research and development in this topic area. Many

current approaches seek to model the attacker and defender by applying existing research in game theory.

(U) Ongoing and future AML and A2I research efforts will seek to identify new attacks and defenses that are not listed above while also discovering methods to improve, optimize, or extend upon those listed above. Current defenses are not measurable and their successfulness usually unquantified. New attacks are being identified on a reoccurring and frequent basis. Applied research and development efforts will react to discoveries made by the research community and integrate those discoveries into prototyped capabilities and solutions.

(U) For the aforementioned reasons, it is critical, to the future of semi-automated, fully automated, and autonomous DoD capabilities, that we research and develop methods to assess AI/ML technologies for their susceptibility to these unique attacks as well as improve their resiliency and safeguard them against these attacks. Accurate and representative modeling of decision support systems, attackers, and defenders will ensure the completeness and reliability of resulting measurements during the assessment of capabilities leveraging AI and/or ML. Modeling will also ensure measurements of the impact(s) of potential adversarial influence are meaningful and quantifiable.

(U) Objectives & Goals

(U) The WG will hold bi-annual two-day workshops. The workshops will provide a collaborative environment that will solicit active participation from all members and attendees. Prior to each workshop a survey will be distributed to members in order to identify high priority topics in which meaningful progress can be made during the workshop. An introduction and conclusion will be held for each workshop, in which all attendees will be brought together. The introduction will briefly cover all of the selected topics for the workshop. For the majority of the workshop, attendees will be divided up into smaller teams that will deep dive into each identified topic. During the conclusion of the workshop, a member from each of the smaller teams will briefly present what they accomplished to all attendees. A summary report and collection of each smaller team's outputs will be captured, documented, and disseminated appropriately within a report. The authoring and distribution of these reports will be coordinated by the WG Chair and/or Co-Chairs.

(U) Deliverables & Timeframes

(U) Events will be held bi-annually, ideally starting in February. Summary reports shall be due within 30 days of the conclusion of the event.

(U) Anticipated Annual Schedule

Deliverable	Year 2020 Projected Timeframe	
	Session 1	Session 2
Workshop	March /April	September/October
Report	April /May	October /November

(U) Anticipated Annual Schedule

Section III: (U) Formation, Staffing, and Organization

(U) Membership Criteria

(U) Each new member will need to submit a statement of interest and completed survey to the WG Chair or a WG Co-Chair who will then grant or deny membership.

(U) Third-party Participants, such as those from academia or industry, may be invited to appropriate events by the WG Chair or a WG Co-Chair to attend events or assist in preparing reports without requiring membership.

(U) Group Formation, Dependencies, and Dissolution

(U) This WG will comprise of an initial founding team of 5-10 that will aim for a total of 40-50 members by the conclusion of the WG's first kick-off meeting. Whether or not membership will be expanded further at that point will be a decision to be made by the WG Chair and WG Co-Chair(s) once their roles are finalized.

The WG Chair and Co-Chairs will be identified and named by the conclusion of the kick-off meeting or shortly after.

(U) Working Group Roles, Functions, and Duties

(U) Participation and support of this WG is given at will and does not commit any person or organization to provide funding or personnel to support its operation.

(U) Army Futures Command (AFC) Combat Capability Development Center (CCDC) Command, Control, Computers, Communications, Cyber, Intelligence, Surveillance and Reconnaissance (C5ISR) Center's Autonomous Cyber effort plans on providing meeting support, document drafting, editing and distribution as well as other substantive contributions when deemed appropriate.

(U) NSA's Laboratory for Advanced Cybersecurity Research (LACR, R2) plans on providing subject matter expertise, meeting support, document editing and distribution as well as other substantive contributions when deemed appropriate.

(U) AFC CCDC Army Research Laboratory (ARL) Network Science Division (NSD) plans on providing subject matter expertise, meeting support, document editing and distribution as well as other substantive contributions when deemed appropriate.

(U) Statements of Interest and Survey

(U) Each member of the Working Group is required to complete a survey and provide a single paragraph statement of interest. The survey will be maintained and provided by the WG.

Section IV: (U) Rules of Engagement

(U) Decision-Making Methodologies

All decisions making for A2IWG will be handled democratically by tentative co-chairs until the lead chair and co-chairs have been formally identified.

(U) Status Reporting

Status reporting requirements will be determined by the lead chair and co-chairs once they have been formally identified.

(U) Problem/Issue Escalation & Resolution Processes

These processes will be determined by the lead chair and co-chairs once they have been formally identified.

(U) Closure & Working Group Self-Assessment

Closure of the working group and its self-assessment process will be determined by the lead chair and co-chairs once they have been formally identified.

It is anticipated that A2IWG will run until at least FY2024 and dependent upon progress, a longer period of time.

Section V: (U) Attachments

- SUMMARY OF THE 2018 DEPARTMENT OF DEFENSE ARTIFICIAL INTELLIGENCE STRATEGY
- A Taxonomy and Terminology of Adversarial Machine Learning Draft - NISTIR 8269

Section VI: (U) Charter Document History

(U) Document Version History

Version	Date	Description
0.X	16 Oct 2019	Draft versions with founding team members
1.0	6 Feb 2020	Updated Charter with tentative role information. Tyler Shipp
1.1	10 Mar 2020	Updated Charter per Brian Haugh's comments and suggestions. Tyler Shipp

(U) Document Version History