2022-06

AN INTRODUCTION TO FRAMEWORK ADAPTATIONS FOR ADDITIONAL ASSURANCE OF A DEEP NEURAL NETWORK WITHIN NAVAL TEST AND EVALUATION

Lyon, Blake A.

Monterey, California. Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

JOINT APPLIED PROJECT REPORT

AN INTRODUCTION TO FRAMEWORK ADAPTATIONS FOR ADDITIONAL ASSURANCE OF A DEEP NEURAL NETWORK WITHIN NAVAL TEST AND EVALUATION

June 2022

By: **Blake A. Lyon**

Advisor: **Robert F. Mortlock**

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2022	**3. REPORT TYPE AND DATES COVERED** Joint Applied Project Report

4. TITLE AND SUBTITLE AN INTRODUCTION TO FRAMEWORK ADAPTATIONS FOR ADDITIONAL ASSURANCE OF A DEEP NEURAL NETWORK WITHIN NAVAL TEST AND EVALUATION	**5. FUNDING NUMBERS**
6. AUTHOR(S) Blake A. Lyon	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

The complexity of modern warfare has rapidly outmatched the capacity of a human brain to accomplish the required tasks of a defined mission set. Task-shedding mundane tasks would prove immensely beneficial, freeing the warfighter to solve more complex issues; however, most tasks that a human might find menial, and shed-worthy, prove vastly abstract for a computer to solve. Advances in Deep Neural Network technology have demonstrated extensive applications as of late. As DNNs become more capable of accomplishing increasingly complex tasks, and the processors to run those neural nets continue to decrease in size, incorporation of DNN technology into legacy and next-generation aerial Department of Defense platforms has become eminently useful and advantageous. The assimilation of DNN-based systems using traditional testing methods and frameworks to produce artifacts in support of platform certification within Naval Airworthiness, however, proves prohibitive from a cost and time perspective, is not factored for agile development, and would provide an incomplete understanding of the capabilities and limitations of a neural network. The framework presented in this paper provides updated methodologies and considerations for the testing and evaluation and assurance of neural networks in support of the Naval Test and Evaluation process.

14. SUBJECT TERMS DNN, test and evaluation, T&E, unmanned systems, unmanned, UAS, unmanned aerial systems	**15. NUMBER OF PAGES** 79
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

AN INTRODUCTION TO FRAMEWORK ADAPTATIONS FOR ADDITIONAL ASSURANCE OF A DEEP NEURAL NETWORK WITHIN NAVAL TEST AND EVALUATION

Blake A. Lyon, Commander, United States Navy

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN PROGRAM MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
June 2022

Approved by: Robert F. Mortlock
Advisor

Robert F. Mortlock
Academic Associate, Department of Defense Management

THIS PAGE INTENTIONALLY LEFT BLANK

AN INTRODUCTION TO FRAMEWORK ADAPTATIONS FOR ADDITIONAL ASSURANCE OF A DEEP NEURAL NETWORK WITHIN NAVAL TEST AND EVALUATION

ABSTRACT

The complexity of modern warfare has rapidly outmatched the capacity of a human brain to accomplish the required tasks of a defined mission set. Task-shedding mundane tasks would prove immensely beneficial, freeing the warfighter to solve more complex issues; however, most tasks that a human might find menial, and shed-worthy, prove vastly abstract for a computer to solve. Advances in Deep Neural Network technology have demonstrated extensive applications as of late. As DNNs become more capable of accomplishing increasingly complex tasks, and the processors to run those neural nets continue to decrease in size, incorporation of DNN technology into legacy and next-generation aerial Department of Defense platforms has become eminently useful and advantageous. The assimilation of DNN-based systems using traditional testing methods and frameworks to produce artifacts in support of platform certification within Naval Airworthiness, however, proves prohibitive from a cost and time perspective, is not factored for agile development, and would provide an incomplete understanding of the capabilities and limitations of a neural network. The framework presented in this paper provides updated methodologies and considerations for the testing and evaluation and assurance of neural networks in support of the Naval Test and Evaluation process.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

A3R	Autonomous Air-to-Air Refueling
AAF	Adaptive Acquisition Framework
ADS	Automated Driving System
AFIT	Air Force Institute of Technology
AI	Artificial Intelligence
AP	Average Precision
AROS	Air Refueling Operator Station
AVO	Air Vehicle Operator
CBTE	Capabilities Based Test and Evaluation
CONEMP	Concept of Employment
COA	Corridor of Autonomy
C4ISR	Command, Control, Communications, Computers, and Intelligence, Surveillance, and Reconnaissance
DARPA	Defense Advanced Research Projects Agency
DAG	Defense Acquisition Guidebook
DAU	Defense Acquisition University
DAS	Defense Acquisition System
DNN	Deep Neural Network
DOD	Department of Defense
DoDI	Department of Defense Instruction
DOE	Design of Experiments
DON	Department of the Navy
DOT	Department of Transportation
DT&E	Developmental Test and Evaluation
EASA	European Aviation Safety Agency
FN	False Negative
FOV	Field of View
FP	False Positive
FRS	Fleet Replacement Squadron
IAW	In Accordance With

IFC	Interim Flight Clearance
IOU	Intersection-Over-Union
ITT	Integrated Test Team
JCID	Joint Capabilities Integration and Development System
LFTE	Live Fire Test and Evaluation
LVC	Live-Virtual-Constructive
MAE	Mean Absolute Error
mAP	Mean Absolute Precision
MIOU	Mean Intersection-Over-Union
MOE	Measure of Effectiveness
MOS	Measure of Success
NAE	Naval Aviation Enterprise
NASA	National Aeronautics and Space Administration
NATOPS	Naval Air Training and Operating Procedures Standardization
NAWDC	Naval Aviation Warfighting Development Center
NATIP	Naval Aviation Technical Information Product
NAVAIR	Naval Air Systems Command
NAVAIRINST	NAVAIR Instruction
NHSTA	National Highway Traffic Safety Administration
NN	Neural Network
NSF	National Science Foundation
NTTP	Navy Tactics, Techniques, and Procedures
ODD	Operational Design Domain
OPTEVFOR	Operational Test and Evaluation Force
OSA	Open Systems Architecture
OT	Operational Test
OT&E	Operational Test and Evaluation
PaRot	Practical Robust Training
PFC	Permanent Flight Clearance
PM	Program Manager
PPBE	Planning, Programming, Budgeting, and Execution
PPM	Project Planning Memorandums

RTA	Run-Time Assurance
RTB	Return to Base
SBTE	Specifications Based Test and Evaluation
SME	Subject Matter Expert
STP	Surrogate Test Platforms
SUT	System Under Test
T&E	Test and Evaluation
TEMG	Test and Evaluation Management Guide
TEMP	Test and Evaluation Master Plan
TIDE	Toolkit for Identifying Detection
TN	True Negative
TP	True Positive
TPID	Test Project Introduction Document
TRMC	Test Resource Management Center
UAS	Unmanned Aerial System
USAF	United States Air Force
USMC	United States Marine Corps
USN	United States Navy
USNA	United States Naval Academy
VRSG	Virtual Reality Scene Generator
XAI	eXplainable Artificial Intelligence

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

As machine learning applications become more capable of performing safety-critical functions, the Department of the Navy will begin to integrate them into manned and unmanned platforms. With the current Department of Navy (DON) Test and Evaluation (T&E) framework though, the machine learning algorithms that the DON attempts to test will not have the proper artifacts to show the algorithms will reliably execute their task in a safe and ethical manner. This research introduces the fundamentals of DON airworthiness and the current T&E framework before reviewing a surrogate test program to examine how the current DON T&E framework could be adapted to provide additional assurance for the fielding of machine learning algorithms.

This research provides stakeholders within the naval aviation enterprise (NAE) an idea of the current tools, methodologies, and frameworks that exist or that are being developed, which, if adapted, could provide additional learning assurance for systems using machine learning to achieve a task currently reserved for a human operator. The primary purpose of the paper has succeeded in that the basics were captured, and, will be verified by actual tests within the coming year. Verification will be accomplished by examining the use case, specifically that the autonomous air-to-air refueling (A3R) of an unmanned aerial system (UAS) with a manned platform, and by relating how new capabilities in the field of assurance could be applied to the test planning, execution, and analysis process. The secondary objectives of capturing the processes by which the initial Test and Evaluation was accomplished, will make the procedures available and extensible to future mission-sets which might be suitable for autonomous systems. As the initial test and evaluation of this particular use-case proceeds, more will be learned, and, when captured, will help to build the foundations for all future naval test and evaluation of autonomous platforms.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to thank a friend and colleague Jonathon K. Parry. Jonathon is currently a doctoral candidate at Purdue Polytechnic University and is the “Deep Neural Net” behind this project.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

This research was developed and written in collaboration with J. Parry for a forthcoming dissertation [1].[1]

With deep neural networks (DNN) becoming increasingly more capable of performing tasks traditionally reserved for humans, systems within Naval Aviation will continue to see an increase in machine learning applications being introduced into legacy, current, and future platforms. A platform is broadly any aerial vehicle. Test and evaluation of a new system or platform serves as an assessment of the ability to safely attain, sustain, and terminate a flight in accordance with (IAW) the prescribed usage and limits [2]. To successfully field new weapons systems reliant on machine learning, though, the existing framework to authorize a permanent flight clearance (PFC) must be adapted to account for the strengths and weaknesses of neural networks. Department of Defense (DOD) instructions are built on the capability to test within a defined domain and rely on a qualified human to operate within the prescribed usage and limits. Given the near-infinite input space available to a neural network, and the possibility that a human may not be able to monitor in or on the loop, instructions must be adapted to account for the unique characteristics of a neural network. Failure to do so will result in an incomplete assurance of the system's performance and will come with a large program cost and longer than the desired schedule; not to mention, it might present a hazard.

As of the date of this publication, no adaptation to the existing DOD airworthiness policy exists that outlines specific tools, methodologies, or framework available to modern-day testers to provide artifacts in support of learning assurance of a neural network. This research provides a basis to structure the approaches for the testing of autonomous systems, within a limited scope. If the DOD desires to maintain parity with our adversary and invest in modern technologies, such as neural networks, the existing airworthiness processes must

[1] Dissertation committee includes: Donald H. Costello, United States Naval Academy, and Joseph B. Sobieralski, Purdue University

update proportionately. The methodologies included herein provide a foremost baseline with a naval context.

The significance of this problem cannot be understated. In 2021, the National Security Commission on Artificial Intelligence published its 756-page final report which provided "an integrated national strategy to reorganize the government, reorient the nation, and rally out closet allies and partners to defend and compete in the coming era of AI-accelerate competition and conflict" [3]. The President of the United States, in the *Interim National Security Strategic Guidance,* stated that "the world's leading powers are racing to develop and deploy emerging technologies, such as artificial intelligence and quantum computing, that could shape everything from the economic and military balance among states to the future work, wealth, and inequality within them" [4]. To be successful in implementing the national strategy, learning assurance must be provided that these systems will operate as intended or true acceptance of this technology will never occur or be increasingly taxing on time or resources. More importantly, failure to develop a family of strategies will put our National Defense at a deficit of preparedness when compared to adversaries.

The primary contributions of this research will comprise introducing stakeholders within the naval aviation enterprise (NAE) to the current tools, methodologies, and frameworks that exist or that are being developed, which, if adapted, could provide additional learning assurance for systems using machine learning to achieve a task currently reserved for a human operator. This will be accomplished by examining a use case, specifically the autonomous air-to-air refueling (A3R) of an unmanned aerial system (UAS) with a manned platform, and by relating how new capabilities in the field of assurance could be applied to the test planning, execution, and analysis process. The secondary objectives are to begin to capture the processes used and make them extensible to future mission-sets suitable for autonomous systems.

The recommendations following this research will be to start educating the NAE's workforce on the findings of this research and pursue additional research. After identifying the tools, methodologies, and frameworks that show promise for adaptation into existing frameworks, collaborative testing between Purdue and The United States Naval Academy

will occur to demonstrate the test planning, execution, and analysis process with these new tools.

The paper is structured as follows. In Chapter II, a background will be provided on current instructions and policies of DOD Airworthiness, focused on Naval Airworthiness, civilian aviation and industry standards, Navy test planning instructions, and ongoing work in the field. Chapter III will provide a brief introduction to Naval Airworthiness and Chapter IV will introduce test planning, execution, and analysis principles. Chapter V will introduce the test case of A3R, while Chapter VI will document the testing process. Finally, Chapter VII will provide conclusions and future research opportunities.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND

This research was developed and written in collaboration with J. Parry for a forthcoming dissertation [1].

A. AIRWORTHINESS

The literature review will be broken-down into multiple sections. First, the current standards available for reference will be presented. Next, current test planning documents will be provided, prior to covering the current frameworks and verification methodologies. Finally, ongoing standards work will be introduced. The standards work, including this thesis, is the basis of a future, distinct documentation set that provides guidance to the T&E workforce on the limited scope testing on UASs.

Airworthiness is defined as "the ability to safely attain, sustain, and terminate flight IAW prescribed usage and limits" [2]. Multiple instructions exist that govern airworthiness certification within the DOD, as documented in [5] to include:

1. DoDD 5030.61: DOD Airworthiness Policy
2. NAVAIRINST 13034.1F: Airworthiness and Cybersecurity Safety Policies for Air Vehicles and Aircraft Systems
3. NAVAIR M-13031.1: NAVAIR Airworthiness and CYBERSAFE Process Manual
4. AFI 62–601: USAF Airworthiness Instruction
5. AFPD 62–6: USAF Airworthiness Policy
6. AR 70-62: Airworthiness of Aircraft Systems
7. MIL-HDBK-516C: Airworthiness Certification Criteria Guidance for the DOD

Additionally, the following military aviation, civilian aviation, and industry standards govern system safety and development assurance, as documented in [5]:

1. MIL-STD-882E: DOD Standard Practice: System Safety

2. Joint Software Systems Safety Engineering Handbook

3. FAA Advisory Circular 23.1309E: FAA System Safety Analysis and Assessment for Part 23 Airplanes

4. European Union Aviation Safety Agency (EASA) Certification Specification 25 Alternate Means of Compliance AMC 25.1309 (Systems and Equipment) is part of the EASA

5. SAE ARP 4761: Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment

6. SAE ARP 4754: Guidelines for Development of Civil Aircraft and Systems

7. DO-178C: Software Considerations in Airborne Systems and Equipment Certification

Collectively, these documents set a foundation to build upon for standards to consider future UASs airworthy.

B. TEST PLANNING

The planning of test and evaluation activities is not a simple or novel task. Each program must develop unique and innovative methods to accomplish its goals; even more so in the case of evaluating NNs. The DOD attempts to capture many of the proven test and valuation (T&E) activities and provide them to program managers (PM) as a guide to success. The guidance provided to PMs is not definitive, but rather a collection of tools to concentrate efforts and streamline processes. The Test Planning Handbook serves as the primary reference for "developing and drafting test project introduction document (TPID), project planning memoranda (PPM), and test plans for tests covered by NAVAIR Instruction (NAVAIRINST) 3960.4" [6]. In support of the Test Planning Handbook, the Defense Acquisition University (DAU) publishes its Defense Acquisition Guidebook (DAG) and the Test and Evaluation Management Guide to provide additional considerations and consolidation of DOD guidance when authoring test planning activities.

After a short discussion of background material, specifically the documentation that is typically used to support airworthiness and general NAVAIR test planning activities, we move on to NAVAIR airworthiness fundamentals.

THIS PAGE INTENTIONALLY LEFT BLANK

III. FUNDAMENTALS OF NAVY AIRWORTHINESS

This research was developed and written in collaboration with J. Parry for a forthcoming dissertation [1]

The Navy airworthiness process is an engineering-based risk mitigation process. It is designed to enable naval leadership to make informed decisions. In the USN, Commander, Naval Air Systems Command (NAVAIR), also known as AIR-00, has ultimate responsibility for both USN and United States Marine Corps (USMC) aircraft [7]. All manned and unmanned aircraft, owned or leased, by the USN and USMC must have airworthiness approval from AIR-00 in the form of a PFC. There are two forms of flight clearances, though, and to get to a PFC: a test program operates under an interim flight clearance (IFC), which is designed to enable limited operations during either a demonstration, such as the X-47B program (Figure 1), or during developmental test and evaluation (DT&E) and operational test and evaluation (OT&E) of a new system or capability.

Figure 1. X-47 Conducts A3R Using Drogue Method. Source [8].

To complete the IFC process, a new capability starts in the center of a build-up process, visualized in Figure 2. From that point, a restrictive envelope is set and the system is cleared for bench testing. Here, items are evaluated at the component or limited integration level. Following successful bench testing, the new system/capability is cleared

for more extensive integrated testing in a laboratory. Following successful bench testing, a system is given a limited ground test envelope. Following a successful ground test, the system would be cleared for a limited flight test envelope.

Normally the flight test envelope is more extensive than the PFC flight envelope. This is because, when conducting a flight test under an IFC, numerous controls are placed on the envelope to ensure various forms of safety. Controls may range from instrumentation installed on the airframe, to numerous flight test engineers monitoring key flight parameters of the air vehicle from a control room. As an air vehicle operated under a PFC would not have such detailed controls available to it, the PFC envelope is more restrictive than the flight test envelope.

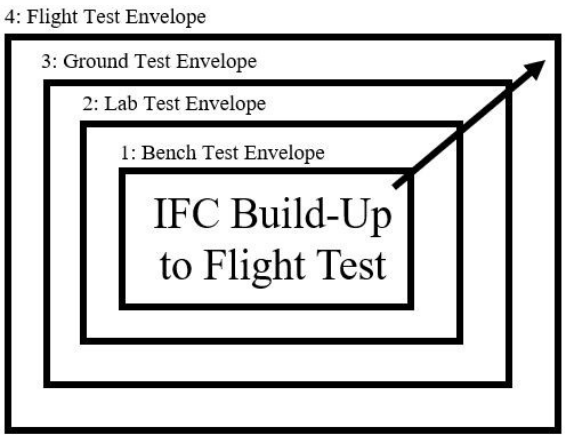

Figure 2. Simplified Example of IFC Process

By direction of AIR-00, the Airworthiness Office of the Naval Air Systems Command (formally known as 4.0P) assesses the aircraft's airworthiness and safety of flight suitability, and ensures all risks have been properly identified within the PFC. The PFC comes in the form of the following two products:

1. Naval Air Training and Operating Procedures Standardization (NATOPS): System Descriptions, Aircraft Operating Limits, Emergency Procedures, and Normal Procedures.

2. Naval Aviation Technical Information Product (NATIP): Armament Systems, Mission Avionics, Store Limitations, Employment Data, and CYBERSAFE program.

Within the USN Airworthiness program, six pillars exist that represent the foundation of the program attributes (Figure 3).

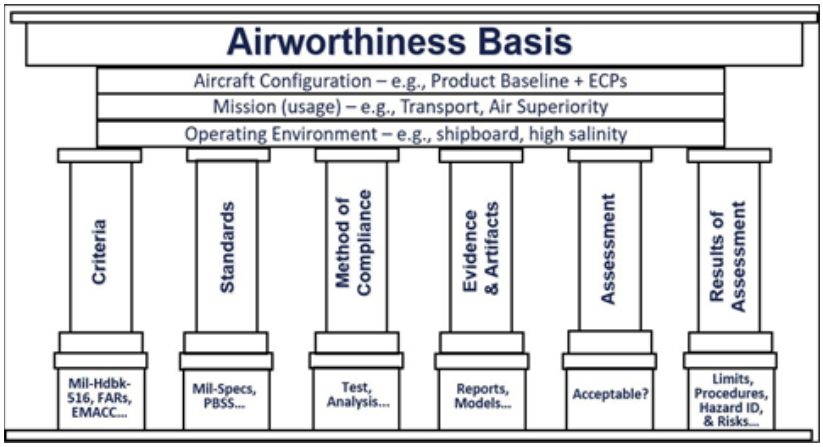

Figure 3. Airworthiness Basics. Source [2].

Those pillars, which aid in a ruling of airworthiness, are defined as follows:

1. Criteria: designed to correct criteria (e.g., MIL-HDBK-516).
2. Standards: meets the standards outlined by the applicable Mil-Specs (e.g., MIL-STD-882E).
3. Method of compliance: evaluation method (e.g., model, flight test, etc.)
4. Evidence Artifacts: specific engineering data.
5. Assessment: SME assessment against tailored airworthiness requirements.
6. Results: non-compliance identified, adjudicated, mitigated, or risks accepted.

While T&E should be included in every step of the airworthiness process, for purpose of this research, the focus will be on the test and evaluation portion (Figure 4).

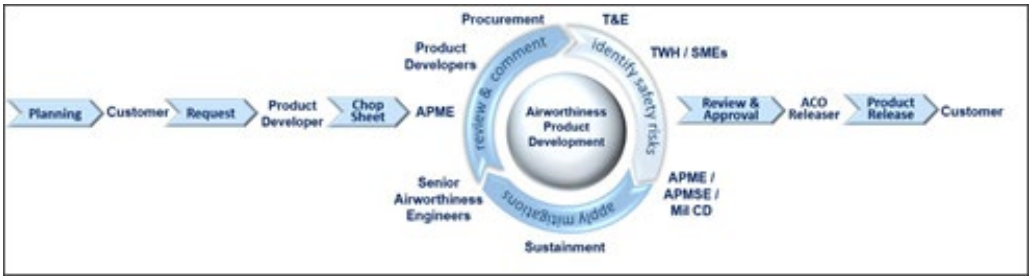

Figure 4. Airworthiness Process

The Naval airworthiness process is also an iterative process (Figure 5), in which sequential progress and feedback codify a platform's readiness.

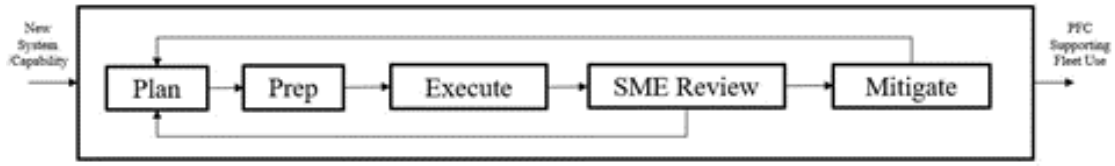

Figure 5. Simplified Iterative Process of Airworthiness

For each new system/platform that enters the airworthiness process, the various phases of the process can be described as follows:

1. Plan: Determine what the configuration of the system under test will be. Additionally, what level of an IFC will this iteration of the cycle support (bench, lab, ground, or flight test).

2. Prep: What is being evaluated and how? In this phase, the various standards, criteria, and methods of compliance for the various technology domains are identified. The actual test matrix (the data that is planned to be gathered during the evaluation of the system under test) is developed.

3. Gather: In this phase, the test matrix, developed in the prep phase, is executed.

4. Subject matter expert (SME) review: In this phase, the data generated during the Gather phase is evaluated by subject matter experts from the

various technology domains against the methods of standards, criteria, and methods of compliance identified during the preparation phase. Additionally, any unforeseen revelations from the data are evaluated by the SMEs. Depending on the results from this phase, more data may be requested and the system/platform would restart the iteration to get more data.

5. Mitigate: In this phase, hazards are identified and gaps are formed from the SME review. There will then be a risk decision made to determine if the next iteration of the process can be executed (e.g., from ground test to flight test).

Each iteration of the airworthiness process is used to support a separate flight clearance. Using the simplified model shown in Figure 5, the initial iteration would support the bench test IFC. The next iteration would support the lab test IFC. The following iteration would support the ground test IFC. The next iteration would support the flight test IFC. Finally, the last iteration would support a PFC for the fleet use of the system. The system is now “airworthy” and can be deployed for real-world operation.

The past chapter summarized the fundamentals of the airworthiness process, to include the engineering-based risk mitigation process, the generalized build-up and iterative cycles, and the foundations of airworthiness basics and summary phases. Next, a short introduction to the structure of naval test and evaluation, as described in terms of organizational structure, policies, and roles and responsibilities.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. FUNDAMENTALS OF NAVY TEST AND EVALUATION

This research was developed and written in collaboration with J. Parry for a forthcoming dissertation [1].

A. THE STRUCTURE OF NAVAL TEST AND EVALUATION

The Department of Defense sets test and evaluation standards for all military departments through the promulgation of various directives, instructions, and memoranda. The test and evaluation of any government procurement effort is mandated by the Defense Acquisition System (DAS) in the Department of Defense Instruction (DoDI) 5000.01 and the Operation of the Adaptive Acquisition Framework (AAF) DoDI 5000.02, and takes various forms, according to strategies defined by the leadership [9]. General T&E guidance, including policy, roles and responsibilities, and procedures are delineated in DoDI instructions [10] and through empowerment by the hierarchy shown in Figure 6.

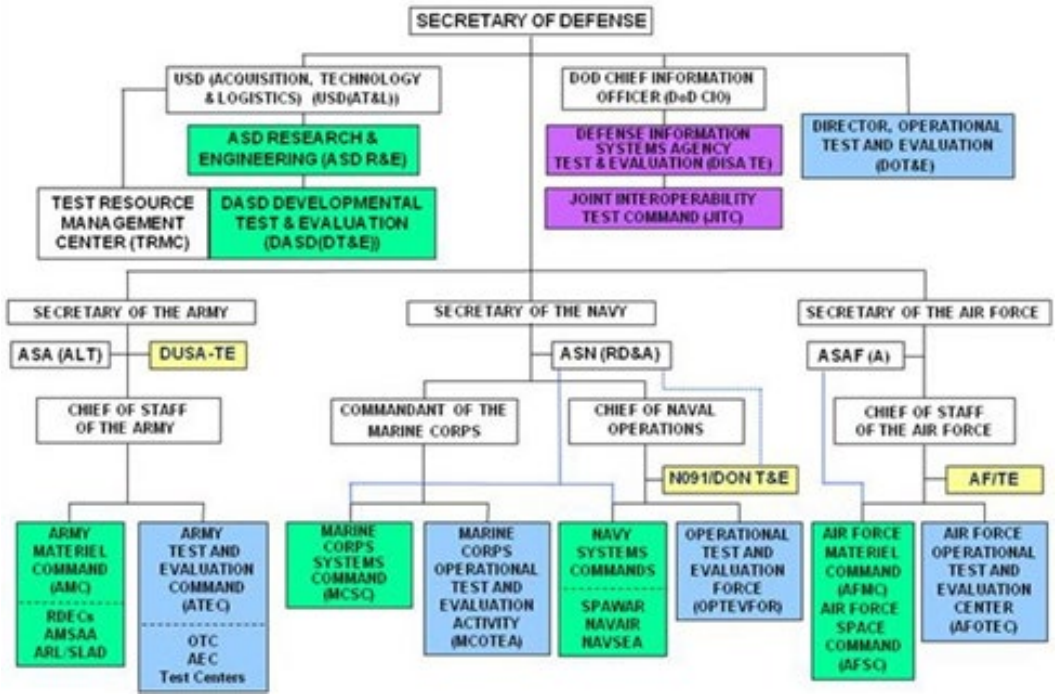

Figure 6. Test and Evaluation Oversight Structure. Source [11].

There are a wide variety of additional guidance sources provided, such as the Defense Acquisition University TEMG [12], in addition to organizational guidance provided to T&E professionals at all levels that meet the specific requirements of each service or specialized unit.

Despite the countless references, instructions, and guidebooks from the various levels of the organizational structure, there is not a large body of official guidance or fit-for-purpose T&E strategies on testing UAS or unmanned systems, especially in the case of platforms that include DNN-enabled capabilities. The DAU TEMG only mentions "unmanned" once in its 313 pages of guidance; never "autonomous" or anything remotely relating to DNNs. The Defense Acquisition Guidebook, Chapter 8–3.21, does indicate that unmanned or autonomous systems have increased considerations for technical complexity for testing, as well as challenges for range and safety approval [13]. This fact applies directly to the T&E efforts which are the convergence of this research, but merely hints at the complications therein.

B. DEVELOPMENTAL TEST AND EVALUATION

DT&E supports data generation for independent organizations, such as the Program Office or Airworthiness, to "measure progress, identify problems, characterize system capabilities and limitations, and manage technical and programmatic risk" [10]. Starting at requirement development, DT will ensure technical requirements are measurable, testable, and achievable to ensure that the appropriate data and assessment will be available to independent decision-makers. The PM and supporting team will develop detailed and integrated test plans for each event outlined in the test and evaluation master plan (TEMP). These TEMPs typically include hundreds to thousands of test points, and involve different types of approaches, such as design of experiments (DOE) or hypothesis testing to claim sufficiency [14]. Most of these tests rely on a small number of preproduction units, executing a waterfall-style buildup approach to the edges of a flight envelope provided by an IFC that initially restricts the flight envelope through an Operational Limit database that can be cleared out through a buildup approach (Figure 7).

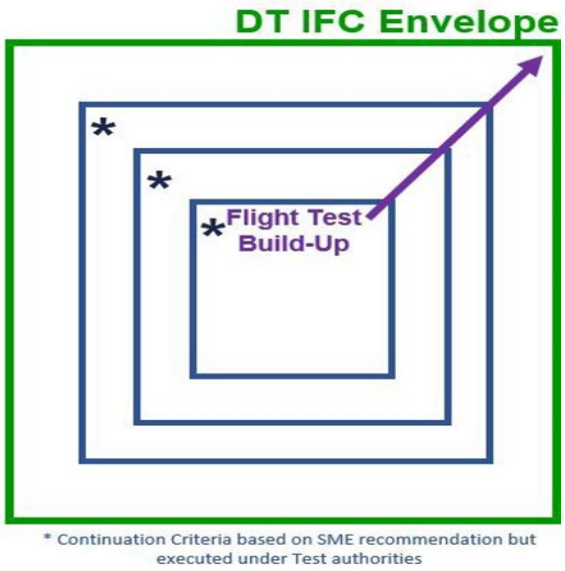

Figure 7. Interim Flight Clearance Buildup Approach. Source [2].

C. OPERATIONAL TEST AND EVALUATION

In the past, operational test (OT) would begin their work following the completion of the entire DT program (Figure 8).

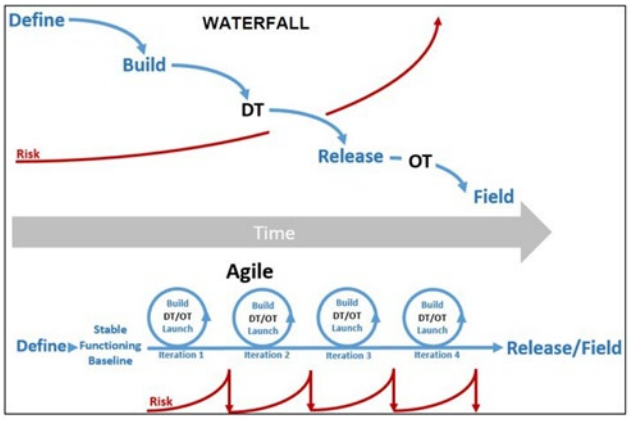

Figure 8. Waterfall versus Agile. Source [15].

Given the pace of the adversary and the unacceptably-long acquisition timelines that the DOD has experienced in the last two decades, this sort of traditional waterfall approach no longer supports the requirements of the fleet and therefore, the DOD has

transitioned some programs to a concurrent DT&E/OT&E approach through an integrated test plan. This system allows OT to test portions of mission sets in the cleared envelope available at the time of test prior to evaluating other portions of the mission once the concurrently expanding, DT established envelope allows. While DT&E will measure progress, identify problems, and characterize capabilities and limitations, OT&E will "estimate a system's military utility, operational effectiveness, and operational suitability as well as the need for any modifications" [16]. Transitioning to integrated T&E does have its advantages as well as disadvantages. Generally, integrated test plans result in a reduction in the total time required for testing, in addition to possible cost savings through a cut of redundant activities. Early involvement of OT&E personnel also gives a chance for early feedback on any test item. Some of the limitations instilled by integrated test include increased coordination requirements, a failure to meet OT or DT test objectives due to more complex or ill-suited test design, and a possible loss of overall system test time as a result of condensed test period.

D. LIVE FIRE TEST AND EVALUATION

The LFT&E portion of the test is mandated by Section 2366 of Title 10, U.S.C. for "all covered systems, munition programs, missile programs, or covered product improvement programs as determined by DOT&E" [10]. These carefully choreographed events involve extensive planning and multiple organizations working together to complete a test point safely. The overall goal of an LFT&E event is to evaluate the survivability and lethality of a system. Although autonomous systems will undoubtedly require LFT&E at some juncture, the premise for this study remains agnostic to the concept. The goal is to first establish airworthiness, then allow some autonomous skills to be established. An autonomous platform must first accomplish the seemingly simplistic goals of basic flight skills—the skills a human might take for granted—prior to being evaluated in realistic combat conditions.

E. CAPABILITIES BASED TEST AND EVALUATION

In recent years, the DOD has adopted the mindset of capabilities-based test and evaluation (CBTE), which moves away from requirements-based performance evaluation,

sometimes referred to as specifications-based test and evaluation (SBTE), and moves toward a broader focus on evaluating a system against a particular operational effect that transcends DT&E, OT&E, and LFTE [15]. This process begins with stakeholders identifying opportunities to use live-virtual-constructive (LVC) evaluations, which involve using a mix of live players, virtual players operated by humans in simulators, and constructive players injected into the environment.

F. THREE PHASES OF TEST

The United States Air Force (USAF) Test Resource Management Center (TRMC) has divided the testing of autonomy and autonomous systems into three categories: test planning, test execution and control, and performance assessment and analysis [15]. For this research, all three phases of test will be introduced, but only the test planning phase will be the focus.

1. Test Planning

Key considerations in the test planning process include data management, design of experiment (DOE), and ensuring there are quantitative mission-focused measures for effectiveness, suitability, and survivability. For the purposes of this paper, we will forgo discussions about survivability as the main focuses are effectiveness and suitability for a certain, limited mission. DOE is the process of designing a statistical methodology for planning, conducting, and analyzing tests and is based on quantitative, mission-focused measures. These measures help to characterize the performance of a system within its mission-based context. Part of the DOE is to detail the desired measures of effectiveness (MOE) of a platform; that is, measure a system's accomplishment of a mission objective or result that is distinctly quantifiable. An MOE might be further decomposed into measures of performance (MOP) and/or measures of suitability (MOS). An MOP is typically a value that is expressed as a number representing a performance feature (e.g., a platform's time-on-station). An MOS represents a system's ability to operate in its environment (e.g., availability). Successful meeting of these measures (MOP and MOS) might indicate a measure of success for a program objective. To evaluate these metrics, a program must accomplish test execution and implement controls.

2. Test Execution and Control

The TRMC further breaks down test execution into four components:

1. Run-time assurance (RTA): The use of a deterministic wrapper that reliably detects a problem and switches to a recovery mode in the event of a failure [15].
2. LVC: Using a mix of live players, virtual players operated by humans in simulators, and constructive players injected into the environment [16].
3. Open systems architecture (OSA): Assuming an RTA has been proven sufficient, using OSA will quickly allow for integration into a proven system.
4. Surrogate test platforms (STP): Using a surrogate testbed, either a fielded manned or unmanned platform, to test new autonomous capabilities.

Each of these four components plays a crucial role in the testing of an autonomous platform and will be combined in the construction of the integrated test plan.

3. Performance Assessment and Analysis

For the purpose of evaluating autonomous platform tasks, MOEs will be analogous to pilot evaluations of platform performance through proven quantitative methods, where the benchmark a system must meet will be a combination of subjective and objective standards [15]. For example, students who conduct aerial refueling for the first time must complete six daytime engagements and six nighttime engagements. As they complete these engagements, there are quantitative methods to apply values to deviations from an expected standard. While an objective minimum exists, a subjective measure of performance will also be provided by the instructor, including general operations around the tanker, engagement attempts, and overall performance. In the same way, instructors must evaluate students during a taking event, both objectively and subjectively, test teams will have to evaluate autonomy in a similar fashion, as machines are limited in ability to explain the decisions and actions that, typically, a human user would make. An objective evaluation of the autonomous vehicle's ability to accomplish tanking being the means and method.

An additional method available, and currently being researched, would be for a machine to have the ability to explain rationale and convey a representation of how it believes it might behave. This would remove the human evaluator from the loop. Work has been done on this subject by the Defense Advanced Research Projects Agency (DARPA) eXplainable Artificial Intelligence (XAI) program using the Army's After-Action Review to organize people's assessment of machine learning in a sequential decision-making environment [17]. However, until this method is proven, humans subjectively evaluating UASs with some quantifiable metrics will be an appropriate surrogate. Now that the different types of test and evaluation were discussed, in addition to test planning, control, and analysis methods, the discussion turns to the mission-related tasks of an autonomous platform. In an effort not to complicate the initial evaluation of UASs, and to present simple and solvable missions, this research has narrowed the tasks to one that will almost certainly be required of any autonomous platform in a naval environment. That mission is tanking.

THIS PAGE INTENTIONALLY LEFT BLANK

V. AUTONOMOUS AIR-TO-AIR REFUELING

This research was developed and written in collaboration with J. Parry for a forthcoming dissertation [1].

Power projection still remains a key component of modern warfare. We must maintain our ability to reach out and touch the adversary. However, projecting power by way of aircraft or warships becomes increasingly difficult as a result of the effective ranges of adversary anti-air weapons and their risks to our aerial forces. In order to reach targets in adversary territory, aircraft must fly correspondingly increasing ranges. In future conflicts, accepting that the UAS is a key component of any mission set, the baseline ranges of what constitutes a long-range mission will surpass the current flight ranges of UAS, and therefore, the ability to conduct A3R will be a critical component of mission success. For the USAF, A3R will be accomplished through the use of a data link system to conduct boom refueling (Figure 9).

Figure 9. KC-46A Boom Refueling an F-15E. Source [18].

Upon receiving a form of "Join Tanking Network" message from the manned tanker, the UAS will establish a data link connection with the manned tanker to exchange messages and control throughout the tanking evolution. The UAS will then proceed to the rendezvous phase with the manned tanker using relative navigation to arrive at the transition point prior to maneuvering to the astern position (Figures 10 and 11).

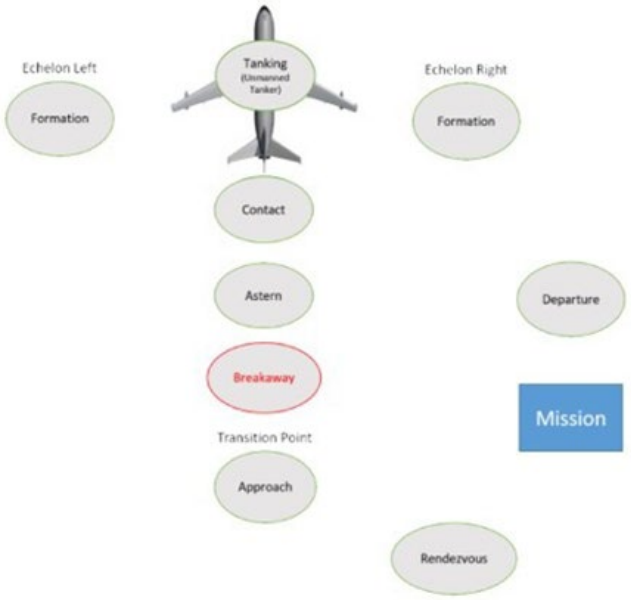

Figure 10. Refueling Phases of Flight. Source [19].

After the UAS moves forward from the astern position to the contact position, a human-in-the-loop in the air refueling operator station (AROS) will take over and complete the contact between the tip of the boom and the aerial refueling receptacle (Figure 11).

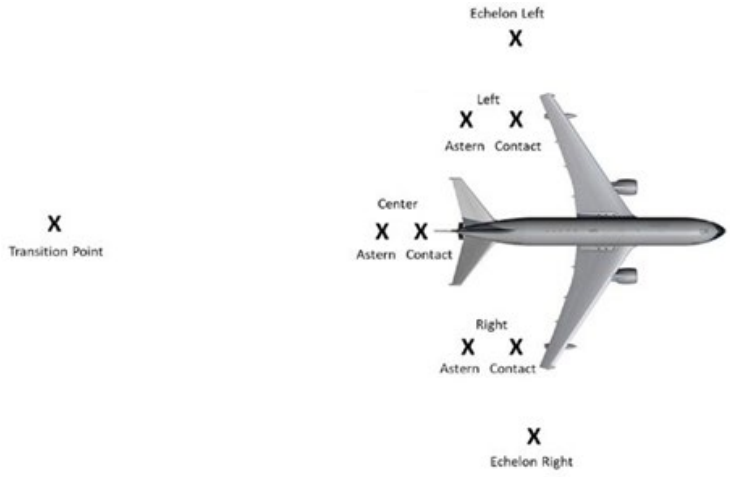

Figure 11. Refueling Positions. Source [19].

While this method works well for the USAF, a UAS operated by the United States Navy (USN) will have to rely on drogue style refueling to extend the range sufficiently to

ensure carrier survivability for the UAS (Figure 1). In drogue refueling, there is no boom operator in the tanker to control the consummation of tanking evolution. All maneuvering and coordination is done by the tanking aircraft.

The tanker aircraft may be fulfilled by a manned (F/A-18E/F Super Hornet) or unmanned (MQ-25A Stingray) platform using a drogue system, since both platforms currently carry the A/A42R-1 refueling store. The A/A42R-1, which is the only fielded refueling store for the Navy, has a 50-foot-long hose that connects to a 27-inch diameter drogue. For the purposes of this test matrix development, the F/A-18E/F Super Hornet was chosen and, therefore, the assumption will be that center position refueling will be conducted. Specifically, the tanking store must be mounted on the aircraft centerline hardpoint. In terms of the concept of employment (CONEMP), the rendezvous and approach shall still be accomplished via data link communication between the UAS and the F/A-18E/F prior to the arrival at the transition point. At the transition point, instead of a boom operator beginning to identify key features at the AROS, a camera system onboard the UAS will begin to use computer vision to develop an understanding of the environment in front of the platform. For purpose of this evaluation, the OAK-D Camera system shall be installed just below the refueling probe tip on the UAS. The OAK-D system, developed by OpenCV, comes equipped with multiple neural networks, to include MobileNet's Single Shot Detector Version 2 (mobileNet-SSDv2), but also has the ability to run neural nets trained in Google's Colab [20].

Upon arriving at the astern position and established in the corridor of autonomy (COA), the camera system, trained on synthetic and live video examples of aerial refueling, will identify key components, such as the drogue, probe tip, and coupler contained within the drogue for the air vehicle operator (AVO) operating the UAS to confirm prior to attempting an engagement. This critical step in the engagement, the identification of the key components of the environment, will serve as the foundation of the test matrix development.

After describing the general tanking evolution with respect to positioning, phases of flight, and required commands, we must now discuss how to test and evaluate this process for an autonomous vehicle.

THIS PAGE INTENTIONALLY LEFT BLANK

VI. TESTING AUTONOMOUS AIR-TO-AIR REFUELING

This research was developed and written in collaboration with J. Parry for a forthcoming dissertation [1].

For this evaluation, the test scenario will be that the DT&E and the OT&E organizations have been tasked to evaluate and provide artifacts in support of obtaining PFC for a UAS to conduct A3R. Specifically, the requirement exists to verify that the UAS is capable of identifying the drogue, the coupler, and the probe tip to an acceptable level to facilitate A3R in mission-representative environments (Figure 12). LabelImg is a common, open-source tool for graphically labeling an image. A product of the image labeling, via LabelImg, in a representative environment is shown in Figure 12.

Figure 12. Labeling Example Using LabelImg. Source [21].

The fundamental concept of this introductory framework is to supplement the existing framework established by the DoDI 5000.89, in that the expectation will be that traditional flight test (flying qualities, loads, noise and vibration, camera evaluation, etc.) will be conducted independent of this framework and that this framework only covers the supplemental considerations and activities required to produce artifacts in support of a PFC associated with the introduction of a computer vision system.

The specific system under test (SUT) will be the DNN operating within the computer vision system, designed and provided by the prime contractor to conduct A3R. As noted previously, the notional MQ-25 will have a PFC for the entire flight envelope with no restrictions for aerial refueling, only to support the requirement to certify the computer vision system to recognize key objects within the domain. A certified RTA exists and will be used, and only the DNN will be tested during this evolution. An F/A-18F Super Hornet (Figure 13) will be the desired surrogate tanking platform for A3R, with the MQ-25 using the A/A42R-1 refueling store (Figure 14).

Assumptions made during the planning process include:

- A proven RTA exists within the MQ-25 architecture. The testing and safety assurance of an RTA will be covered in future research.
- An STP is available and for purposes of this research, a F/A-18F Super Hornet will serve as the STP.
- Test Instrumentation exists which is capable of capturing near real-time data of the computer vision system. A list of instrumentation requirements will be provided later in the research.
- No flight restrictions will inhibit the testing of either the UAS or STP in all required weather conditions suitable for normal refueling
- This program will supplement the existing framework of test and evaluation and therefore, discussion over the camera, vibration concerns, flying qualities in the pre-contact precision, or similar topics will not be covered in this research.

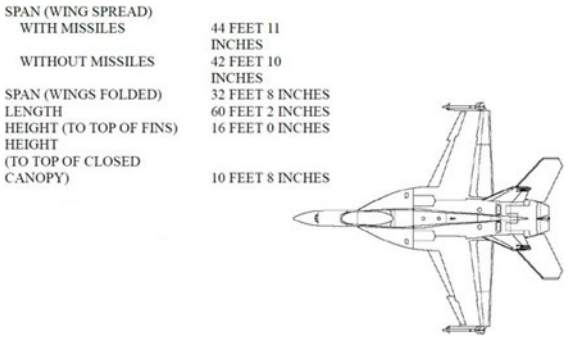

Figure 13. F/A-18 Super Hornet: Surrogate Test and Tanking Platform. Source [22].

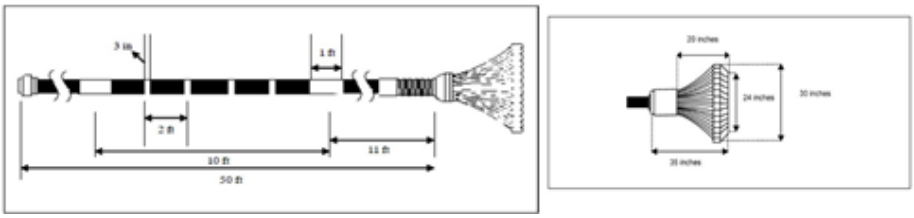

Figure 14. A/A42R-1 Refueling Store Hose Assembly. Source [22].

A. TRAINING LIMITATIONS

When developing a test strategy to provide artifacts of a DNN, it is critical to understand that exhaustively testing a DNN will not be feasible, and gaining 100 percent assurance that the DNN will correctly perform as desired will not be attainable [23]. Current software assurance frameworks fundamentally rely on the fact that only a finite set of potential hazards/failures may be identified beforehand and will not change over the life of the software [24]. However, for machine learning applications, such as computer vision, the input space in the operational domain approaches infinity, and therefore, defining all possible hazards or failure states will not be achievable.

Even a network demonstrating high levels of accuracy on test data may end up failing to function as designed in the face of an adversarial attack [25]. In the specific use of A3R, the risk of an adversarial attack in the operating domain are low given the proximity of the receiver to the tanker and the proximity to the refueling evolution required to affect the 20 feet of the operational domain. For other applications relying on machine learning where the adversary may manipulate the environment, such as target recognition,

adversarial attacks must be considered and assurance must be provided that the DNN will not succumb to those attacks, as in Figure 15 [26]. Figure 15 is an example of how an image, at the macro level, might appear to be entirely unperturbed to the human eye (right, adversarial image). However, because of slight variations made to the image at the micro level, a DNN trained to accomplish target recognition could be misled.

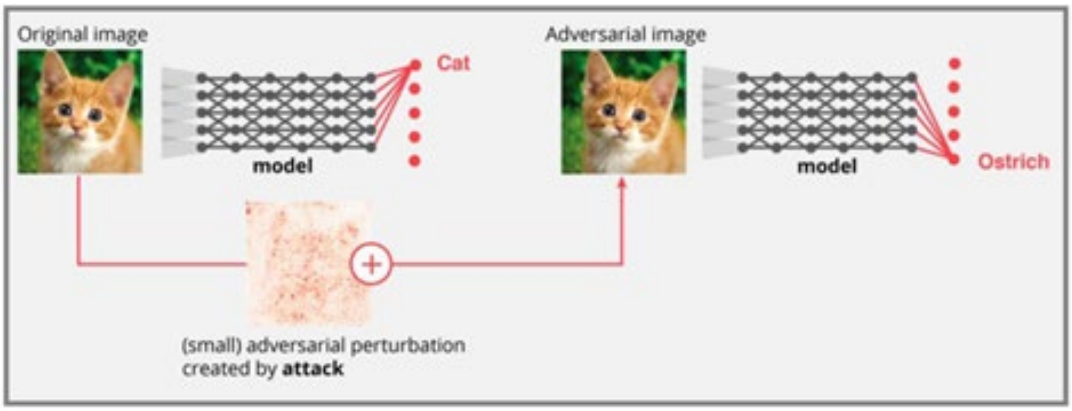

Figure 15. Notional Example of an Adversarial Attack

B. FRAMEWORK REQUIRED PRIOR TO TEST

To successfully test a neural network to produce the desired artifacts, integration of the test team with the prime contractor must occur as early as possible. Given the depth of required knowledge to effectively and efficiently test a DNN, a tiger team should be assembled consisting of both military and civilian members of DT&E (both Air Vehicle and Mission Systems), OT&E, platform-specific members from the Naval Aviation Warfighting Development Center (NAWDC), and stakeholder representatives from the program office(s) concerned. This approximately ten-person team will integrate with the prime contractor's development team to create an integrated test team (ITT), charged with producing the artifacts required to support a PFC in an operationally representative environment. A summary of the machine learning development cycle provided by the European Aviation Safety Agency (EASA) overlaid with acquisition milestones is provided in Figure 16 as a baseline for activities [5].

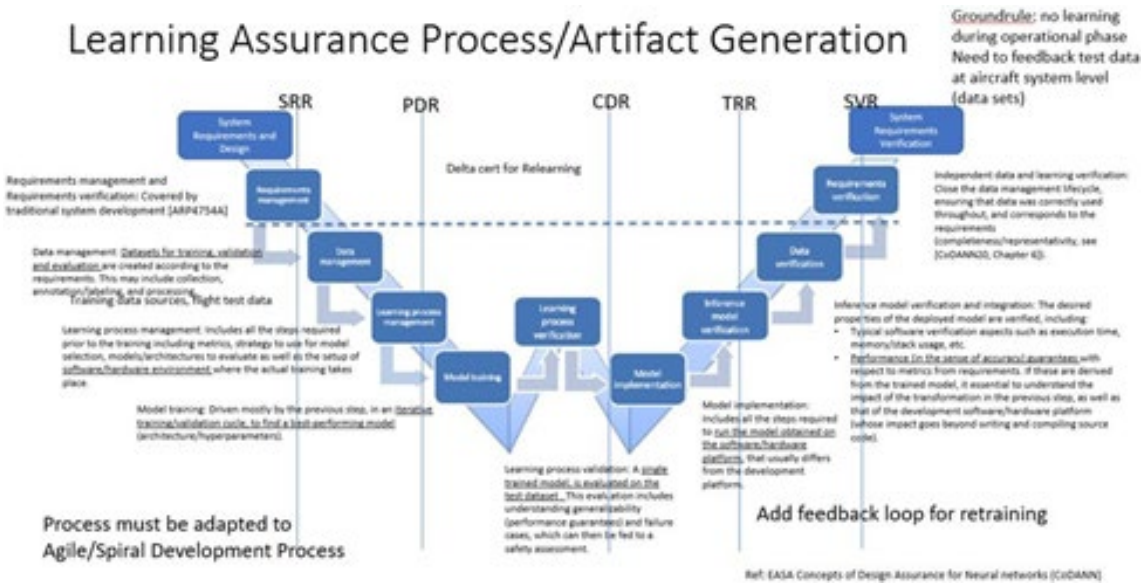

Figure 16. CoDANN W Overlaid with Acquisition Milestones. Source [5].

When training and evaluating a DNN, a representative and comprehensive environment must be defined to ensure that assurance throughout the entire desired flight envelope is achieved. The problem with developing this is that current approaches to certification assume there are a finite set of potential hazards/failures that may be identified beforehand, and will not change over the lifetime of the system [24]. One method created by National Highway Traffic Safety Administration (NHSTA) to methodically define the operating domains in which automated driving systems (ADS) are operational design domains (ODD) [27].

Using a similar approach, an ODD for the task of A3R was created to document the different possible domains that a UAS may operate in to accomplish the task of A3R. The NHSTA defines the following six top-level categories to consider when designing an ODD:

1. Physical Infrastructure
2. Operational Constraints
3. Objects
4. Connectivity
5. Environmental Conditions

6. Zones

For an ADS, the domain may include roadway type, speed, lighting, weather, objects, connectivity, and zones. From there, a hierarchical approach may be constructed to test the different domains that ADS may operate in (Figure 17).

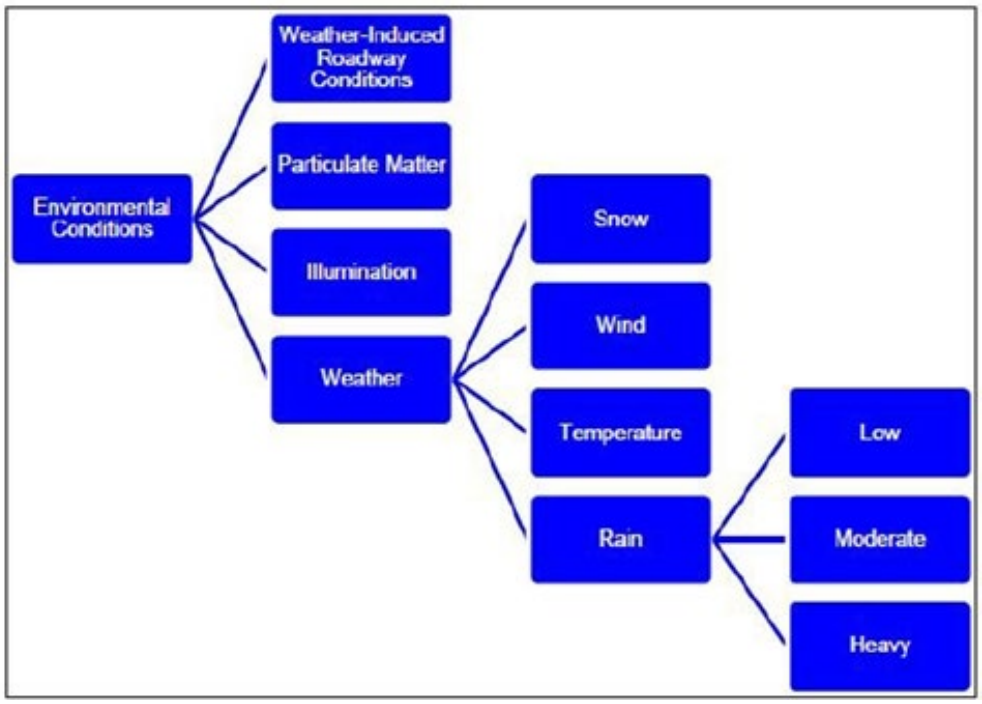

Figure 17. Example Hierarchical Levels in Environmental Conditions for ADS. Source [27].

Using a similar approach, a review of the intended mission sets provided by the tiger team will inform the ODD taxonomy. An example ODD for the A3R outlines examples of the key categories involved (Figure 18).

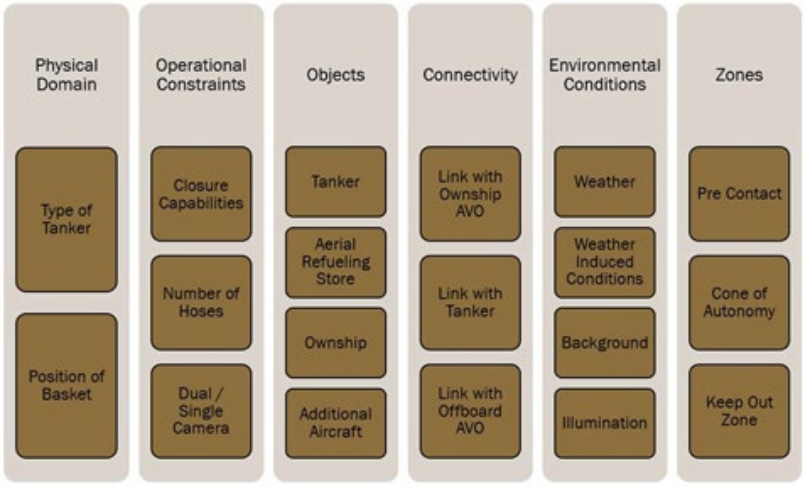

Figure 18. Draft Operational Design Domain for A3R

Each top-level category may then further be broken down into specific domain characteristics, as shown for environmental conditions (Figure 19).

When designing the altitude component of the domain, coordination must be done with the receiver certification team to ensure as much overlap has been created between the ODD and the receiver certification test plan. An example section of a notional receiver test plan would include configuration, receiver loadout, initial airspeed (KCAS), initial altitude (K ft. MSL), and remarks (Figure 20) [28].

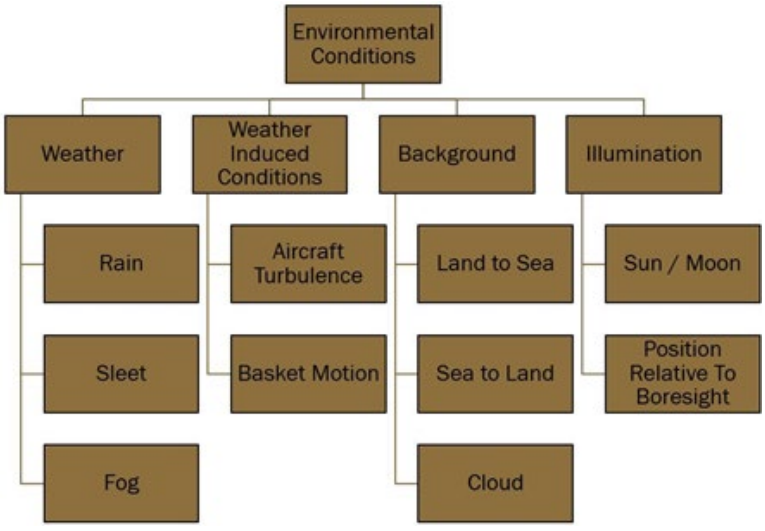

Figure 19. Draft Operational Design Domain for A3R – Environmental Level

Test Point #	Maneuver	Receiver Configuration	Receiver Loadout	Initial Airspeed (KCAS)[2]	Initial Altitude (K ft MSL)[3]	Remarks[1]
0100	Pre-Contact Wake Survey	Cruise	Clean	260	15	-
0110	Contact/Disconnect – S&L	Cruise	Clean	260	15	-
0120	Drogue Tracking – Turn	Cruise	Clean	260	15	-
0130	Contact/Disconnect – Turn	Cruise	Clean	260	15	-
0200	Pre-Contact Wake Survey	Cruise	Clean	180	20	-
0210	Contact/Disconnect – S&L	Cruise	Clean	180	20	-
0220	Drogue Tracking – Turn	Cruise	Clean	180	20	-
0230	Contact/Disconnect – Turn	Cruise	Clean	180	20	-
0300	Pre-Contact Wake Survey	Cruise	Clean	260	35	-
0310	Contact/Disconnect – S&L	Cruise	Clean	260	35	-
0320	Drogue Tracking – Turn	Cruise	Clean	260	35	-
0330	Contact/Disconnect – Turn	Cruise	Clean	260	35	-
0400	Pre-Contact Wake Survey	Cruise	Clean	180	5	-
0410	Contact/Disconnect – S&L	Cruise	Clean	180	5	-
0420	Drogue Tracking – Turn	Cruise	Clean	180	5	-
0430	Contact/Disconnect – Turn	Cruise	Clean	180	5	-
0500	Pre-Contact Wake Survey	Cruise	Clean	300	5	-
0510	Contact/Disconnect – S&L	Cruise	Clean	300	5	-
0520	Drogue Tracking – Turn	Cruise	Clean	300	5	-
0530	Contact/Disconnect – Turn	Cruise	Clean	300	5	-
0600	Pre-Contact Wake Survey	Cruise	Clean	300	29.4	-
0610	Contact/Disconnect – S&L	Cruise	Clean	300	29.4	-
0620	Drogue Tracking – Turn	Cruise	Clean	300	29.4	-
0630	Contact/Disconnect – Turn	Cruise	Clean	300	29.4	-
0700	Pre-Contact Wake Survey	Gear Down	Clean	180	20	Perform at pilot discretion
0710	Contact/Disconnect – S&L	Gear Down	Clean	180	20	-
0720	Drogue Tracking – Turn	Gear Down	Clean	180	20	-
0730	Contact/Disconnect – Turn	Gear Down	Clean	180	20	-
0800	Pre-Contact Wake Survey	Gear Down	Clean	230	20	Perform at pilot discretion
0810	Contact/Disconnect – S&L	Gear Down	Clean	230	20	-
0820	Drogue Tracking – Turn	Gear Down	Clean	230	20	-
0830	Contact/Disconnect – Turn	Gear Down	Clean	230	20	-
0900	Pre-Contact Wake Survey	Gear Down	Clean	180	5	Perform at pilot discretion
0910	Contact/Disconnect – S&L	Gear Down	Clean	180	5	-
0920	Drogue Tracking – Turn	Gear Down	Clean	180	5	-
0930	Contact/Disconnect – Turn	Gear Down	Clean	180	5	-
1000	Pre-Contact Wake Survey	Gear Down	Clean	230	5	Perform at pilot discretion
1010	Contact/Disconnect – S&L	Gear Down	Clean	230	5	-
1020	Drogue Tracking – Turn	Gear Down	Clean	230	5	-
1030	Contact/Disconnect – Turn	Gear Down	Clean	230	5	-

Figure 20. A Portion of a Notional Receiver Certification Test Matrix. Source [28].

Before the prime contractor has the ability to train the DNN, data must either be collected using a real-world surrogate aircraft, created using a simulation engine like MetaVR [29], or must be augmented using real-world data [30], to train, test, and evaluate the performance of the DNN. Unless the decision is made to use only synthetic data to train the DNN, the collection of real-world data must begin well prior to the training of the DNN to allow for the preparation to be complete in time to train the DNN. A general flow of preparation of a data set includes the following steps: video collection, image frame

extraction, image frame selection, image labeling, image and label resizing, image filtering, and data augmentation [31]. To collect real-world video, a camera must be installed within a surrogate platform in such a manner that the probe tip, drogue, couple, and other desired features are not obscured or obstructed by the structure of the platform. Alternatively, a camera system may be modified on the exterior of the surrogate platform, but this would require additional risk mitigation for both the test program, due to limiting the number of assets available to record, and the surrogate test platform, due to potential foreign object debris (FOD) or aerodynamic issues.

While the prime contractor completes the remaining steps in preparation of the data set, the tiger team should be involved to ensure an understanding is had among the ITT on the methodology involved to collect, label, and process the data while continuing to provide operational context to the domain (Figure 21) [32].

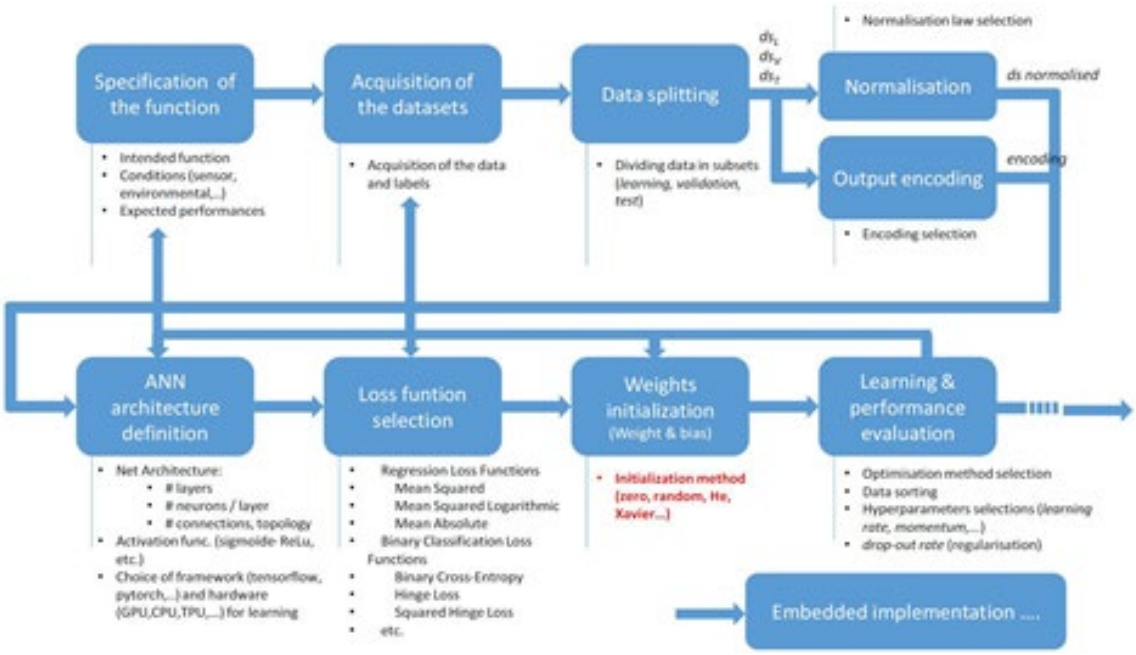

Figure 21. A Simplified Workflow for Supervised Machine Learning. Source [32].

The prime contractor should strive to develop and deliver in an agile structure, delivering incremental capability versus one large delivery. Working with the tiger team,

an understanding should be had among the ITT of the next desired incremental capability and the ODD should reflect the path intended to achieve the full desired capability.

C. PROPOSED FRAMEWORK OF TESTING

Once the DNN has been trained to a sufficient level of performance to release an incremental capability using the training portion of the labeled data set and evaluated against the testing portion of the labeled data set, the DNN may now be handed over to the government test team for evaluation. Since the tiger team has been participating in the design and development of the DNN, a dedicated product outlining the results of the training should not be required to transition, but rather be accomplished concurrently with the start of training for reference later in testing. The cost of dedicating a tiger team to this effort begins to show the benefit with opportunities to rapidly transition the DNN from phase to phase.

The proposed framework for testing breaks down as follows. First, formal methods will be produced in support of further understanding of the overall robustness of the DNN. Next, combined simulator and laboratory testing will be conducted for the government to evaluate the DNN prior to installation on an STP. Depending on the STP, surrogate testing may begin either simultaneously with simulator and laboratory testing (manned platform operating the DNN in the background) or subsequently (unmanned platform or manned platform operating the DNN as a primary means of control). Finally, the DNN may then be installed on the test platform, the MQ-25 for this research, and evaluated in the open-air. Ideally, the formal methods, simulator testing, laboratory testing, and surrogate test platform (background) testing would occur near-simultaneously in an effort to decrease time to provide artifacts for learning assurance. Early on in test, this may not be a possibility, but as additional incremental capabilities of the DNN are developed and delivered, this would become the ideal scenario.

1. Formal Methods

Formal methods are a rigorous mathematical technique used for the development and verification of software or software-based systems. Formal methods ensure these

systems are developed without error to ensure real-world functionality, dependability, and consistency of a product.

In the DOD 5000.89 instruction, DT&E activities are designed to provide data for independent evaluations and include an assessment of the system's capabilities, limitations, and deficiencies [10]. With DNNs, this requirement presents a unique requirement to assess the robustness of the model. As shown in [16], small perturbations in the input x, may lead to drastically different labeling of objects in the domain y. One approach to assessing the susceptibility to perturbations includes approximation approaches or targeted attacks, but these approaches are heuristic-based and only operate on individual input points, which may not accurately document the DNN robustness [33].

Multiple different research efforts have been and are currently being conducted to overcome the issue with local robustness. DeepSafe applies a clustering algorithm over regions of inputs from the training portion of the data set that has a high probability of being labeled consistently. Each cluster then has a centroid and radius defined in which the same label should be assigned to all inputs within that region. By decomposing the requirement to show global robustness to clusters, tools such as Reluplex [34] may be used to efficiently solve [33].

Practical Robust Training (PaRot), a framework introduced by a research team out of Cambridge University based in TensorFlow [35], allows for robust training and testing of existing DNN. They define their work as veritably robust training and as having the ability to take existing graphs within TensorFlow to allow for faster development and testing [36]. This field has grown so large, that an annual international competition is now held as part of the Workshop on Formal Methods for ML-Enabled Autonomous Systems called the International Verification of Neural Networks Competition (VNN-COMP 2021) [37]. This competition consisted of twelve teams with the goal of providing objective comparisons of different methods in terms of scalability and speed and was supported by the USAF, the DARPA, and National Science Foundation (NSF).

In summary, the field of formal methods shows promise for assisting in the activities required by the DOD 5000.89. While only a few methods were discussed in this

research, additional methods may be found in the proceedings from the National Aeronautics and Space Administration (NASA) Formal Methods International Symposium from previous years, including 2021 [38], 2020 [39], 2019 [40], 2018 [41], and 2017 [42]. While every method applicable to computer vision was not covered in this research and no single method was recommended for integration into the NAVAIR T&E process, the goal of this chapter was to introduce the reader to the concept of formal methods, the reason to use formal methods, and examples of formal methods currently available at time of publishing. The tiger team should work with the prime contractor during the training of the neural network to determine the best tools available for the team to use and then integrate those into the testing framework.

2. Simulator Testing

Simulator and Laboratory testing will be a crucial role in providing assurance of a computer vision system tasked to identify key aspects of a tanker. As outlined in the explanation of the ODD, a diverse and extensive domain exists encompassing the different environments that the DNN could encounter in normal flight operations. Testing the entire ODD in the open air will not be possible due to the complexity of coordinating aircraft, reliance on the environment, and time required to complete testing. Therefore, a combination of simulator and laboratory testing must be developed, certified, and then used as a supplement to open-air testing in support of collecting artifacts. This chapter covers the adaptations to existing simulator testing that are required.

Simulation environments exist that could be used to test a DNN for the purposes of object detection, such as MVRsimulation Virtual Reality Scene Generator (VRSG) [29]. These environments are capable of producing the diverse domains required to evaluate the DNN and could be used for both creating synthetic data to train and test the DNN for the prime contractor as well as evaluating the DNN in support of a government evaluation. Metrics must be calculated and collected from the simulation environment and will be introduced in the next chapter.

Before beginning, a key concept to understand is intersection-over-union (IOU). IOU is a key measurement of performance and is represented by the common area between the object detected and the labeled object over the union of the two areas (Figure 22) [43].

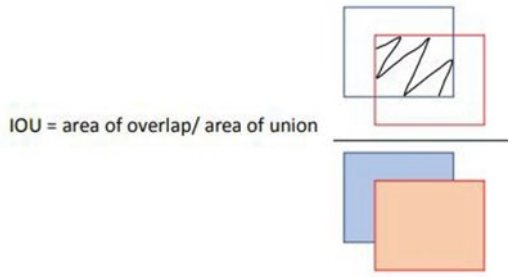

Figure 22. Intersection-Over-Union (IOU)

A low IOU means that the common area between where the DNN detected the object and the actual location of the object is proportionally small. In the A3R use case, a positive detection with an IOU of 0.10 would mean a majority of the detected drogue bounding box does not overlap the actual drogue bounding box. Conversely, an IOU of 0.95 would mean a majority of detected drogue bounding box overlaps the actual drogue bounding box. Therefore, an IOU must be agreed upon prior to calculating metrics such as true positive (TP), true negative (TN), false positive (FP), or false negative (FN). If the decision to make IOU = 1.00 is made, then only TP with 100 percent overlapping detection and actual bounding boxes will be classified as a TP. Therefore, based on the size of the drogue and coupler, a determination must be made on what an acceptable IOU must be to complete an intercept of the probe tip to the coupler. For purpose of this evaluation, an IOU of 0.95 will be set for the calculation of TP, TN, FP, and FN.

Additionally, in that the number of FP and FN relies on the confidence score threshold chosen, and therefore, an appropriate confidence score must be chosen and documented for review by independent reviewers [44].

Having set an appropriate IOU and confidence score, the following equations and definitions are referenced from [45].

$$MIOU = \frac{1}{n}\sum_{i=1}^{n} IOU_i$$

Mean intersection-over-union (MIOU) represents the mean value of IOU.

$$P = \frac{TP}{TP + \text{FP}}$$

Precision (P) represents the percentage of times the DNN correctly identifies the couple, drogue, and probe tip.

$$R = \frac{TP}{TP + FN}$$

Recall (R) represents the percentage of actual objects that were correctly classified.

$$MAE_x = \frac{1}{n}\sum_{i=1}^{n} |\overline{\Delta x_i}|$$

$$MAE_y = \frac{1}{n}\sum_{i=1}^{n} |\overline{\Delta y_i}|$$

Mean absolute error (MAE) represents the average number of pixels in the x and y-axis respectfully of all TP images. This could also be calculated as an angular accuracy. The following equations are referenced from [46].

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n}$$

Mean absolute precision (AP) represents the mean value of the average precision (AP) and is defined by the area under the precision-recall curve. This metric has been used as a primary means of describing the performance of object detection for many years, but it does suffer from some shortcomings. If the team only optimizes for mAP, then the importance of different error types (Figure 23) may be omitted from the final findings [47].

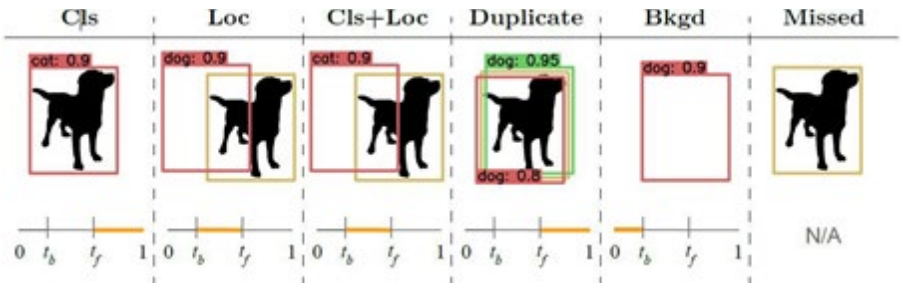

Figure 23. Error Types

The error types are categorized as:

1. Classification Error: Correct localization, but incorrectly classified.
2. Localization Error: Correctly classified, but incorrectly localized.
3. Both Classification and Localization Error: Incorrectly classified and localized.
4. Duplicate Detection Error: A higher scoring detection lead to incorrect classification.
5. Background Error: Background detected as the foreground.
6. Missed Ground Truth Error: All undetected false negatives not covered.

To address the shortcomings of only optimizing for mAP, the Toolkit for Identifying Detection (TIDE) framework was developed to summarize error types, isolate the contribution, compare across data sets, incorporate all predictions, and allow for finer analysis [47].

3. Lab Testing

As simulator testing begins, laboratory testing may also begin to test the performance of the DNN on the actual camera system. To evaluate the performance of DNN in different domains inside the ODD, augmented reality may be used to simulate different environments. Just as augmented reality may be used to create training data [29], augmented reality may be used to assist in the evaluation. Air Force Institute of Technology

(AFIT) has used this concept in recent years to test the USAF method of aerial refueling when the boom occludes the receiving aircraft [48].

As with simulator testing, the laboratory testing will produce metrics such as precision, recall, and mAP, but will also provide additional metrics of performance associated with the camera, such as the temperature, the processing power required, and the frame rate. These metrics fall under traditional hardware testing though and will not be covered in detail in this research.

4. Surrogate Test Platform

Given the requirement for the DNN to function as desired on day one of the receiver certification program, consideration should be given to conducting testing on an STP prior to installing the DNN on the objective platform. An example of this includes the 2006 testing conducted by NASA using a surrogate F/A-18 to conduct autonomous airborne refueling [49] or the X-47B program previously discussed in this paper [8].

This could be accomplished either on a surrogate UAS that has been cleared for flight with an operational RTA or be accomplished in the background of a manned platform during air-to-air refueling. Given the use case of A3R, no platform currently exists, though, that could serve as a surrogate. For other mission sets though, the USAF has demonstrated this concept with a Kratos UTAP-22 tactical unmanned vehicle in 2021 [50], the Kratos produced XQ-58A Valkyrie, and the General Atomics produced MQ-20 Avenger [51]. In the future, once a cleared platform exists to conduct A3R, new DNNs that provide additional domain capabilities may be installed on that platform to serve as a surrogate to test additional capability in a much more rapid test schedule.

In terms of a manned platform, consideration should be given to clearing the desired camera system for installation into current platforms that conduct air-to-air refueling, such as the F/A-18E/F, EA-18G, or F-35C, and allow the system to run in the background during manned flight operations. This comes at no additional cost to the test program (after installation of the camera system) in terms of flight test funding since it would serve as tag-along testing to another program requiring manned air-to-air refueling. Post-flight, the test

team could perform a qualitative assessment of the performance of the tracking system or label data post-flight to measure against the metrics outlined earlier in the paper.

5. Capabilities Based Test and Evaluation of UAS

Aerial refueling testing consists of qualification and certification [28]. Qualification involves showing that the aircraft under test meets specification standards and meets the mission requirements. The certification consists of the testing of two qualified aircraft and verifying functionality and capability as a pair throughout the desired tanking envelope. For this scenario, the assumption is that the tanking aircraft is qualified and the focus of the flight test program will be qualifying and certifying the MQ-25.

A receiver certification test program will include components such as "receiver aircraft performance and handling qualities, receiver pilot visual aids on the tanker and drogue system, and receiver workload to refuel" [28]. To complete this program, maneuvers such as a wake survey (Figure 24) or drogue tracking (Figure 25), must be conducted, and to do this successfully, the DNN must be correctly classifying and localizing the drogue.

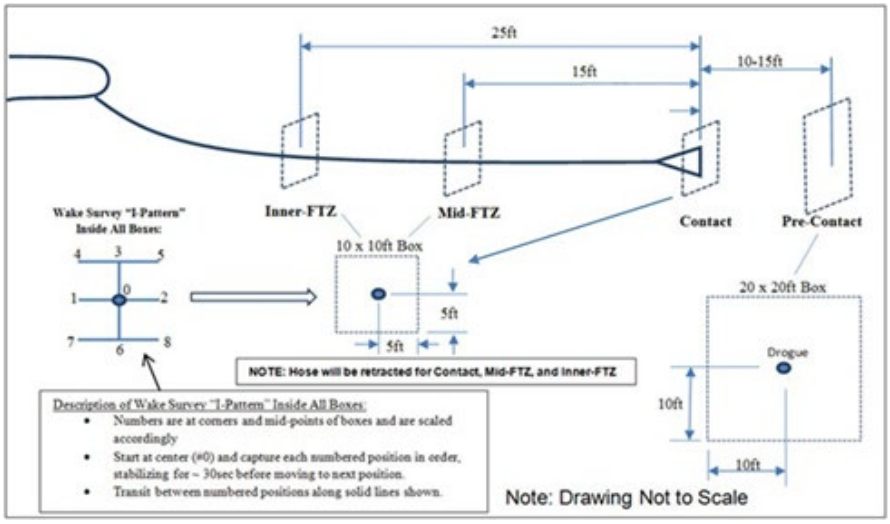

Figure 24. Wake Survey I-Pattern and Receiver Positions. Source [28].

The general buildup of the test program will start with wake surveys, drogue tracking, both straight and level in turns and then straight and level contacts. Understanding the existing framework for a receiver certification test plan and the fact that the DNN must be operational for all points, the assurance test matrix should be categorized as "tag-along" testing with a minimal number of independent test points being added for the sole purpose of producing artifacts in support of assurance.

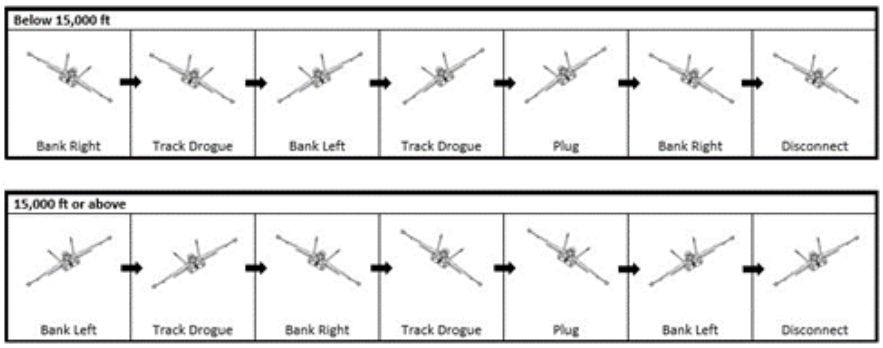

Figure 25. Combined Drogue Tracking and Turn Maneuver

As was described, formal methods provide a rigorous mathematical technique that was used for the development and verification of software or software-based systems. Formal methods ensured that the progression of test phases (i.e., simulator, lab, surrogate) were developed without error to ensure real-world functionality, dependability, and consistency of a product. The results of these events help to provide data for the qualification and certification of the mission-set. Next, we must use the data that was gathered and execute a performance assessment and analysis.

D. PERFORMANCE ASSESSMENT AND ANALYSIS

Both DT&E and OT&E are charged with assessing the performance of the system against a specification and in an operational environment. In traditional flight test, real-time telemetry will allow for monitoring of critical parameters in support of a quantitative assessment whereas the human-in-the-loop (the test aircrew) will provide an additional qualitative assessment of handling qualities (Figure 26) or workload required to complete a task (Figure 27).

Some of the methods used to apply quantitative measurements to typically qualitative assessments are through the use of some common rating scales, such as the Cooper-Harper Handling Qualities Rating Scale (Figure 26). The Cooper-Harper Scale allows the handling qualities of an aircraft to be quantified on a scale of 1–10, in addition to gross, generalized handling characteristics such as having "major deficiencies" in the way the aircraft maneuvers or having qualities that are "excellent" or "highly desirable." Similar handling scales, typically used for manned aircraft, could be used to characterize the flight characteristics of a UAS during mission-sets, such as A3R; just from a different perspective. In addition to flight characteristics, an individual observing a UAS execute a mission could use the Revised Bedford Workload Scale (Figure 27) to quantify the difficulty, or amount of work or capacity, that certain tasks require.

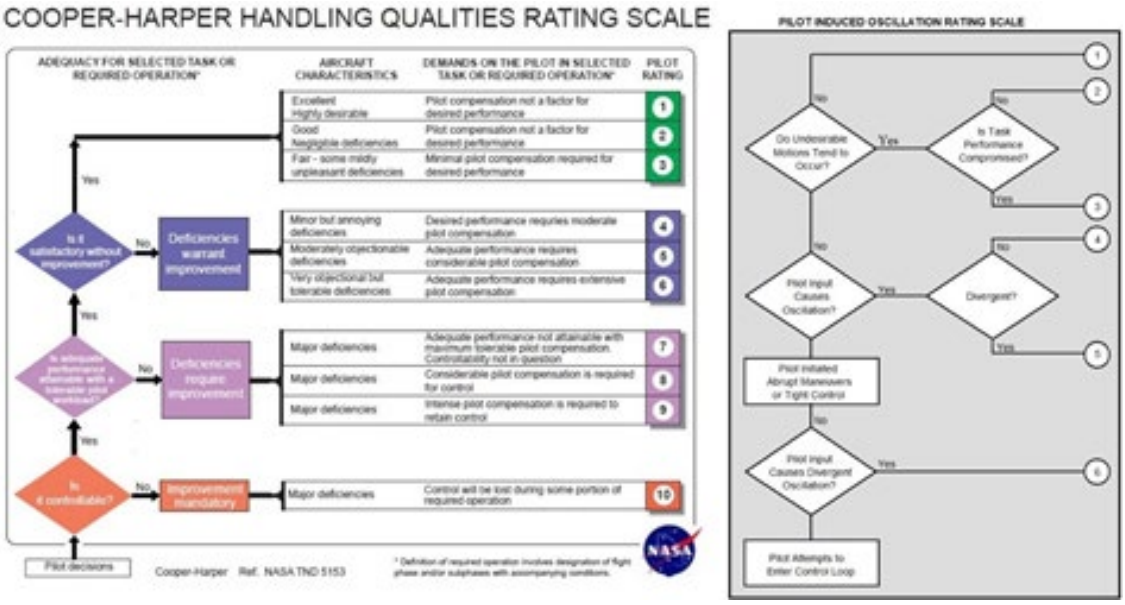

Figure 26. Handling Qualities Rating and Pilot Induced Oscillation Scale. Source [28].

During open-air operations of the test UAS, real-time quantitative monitoring of metrics defined in the simulator and laboratory testing will not be possible. Those metrics all rely on labeled data to compare against the performance of the DNN. Since real-time labeling of the critical objects within the field of view (FOV) of the camera will not be

possible, only qualitative remarks may be made during open-air testing and a new method must be adapted to allow test engineers a way to organize their thoughts and logically think through the actions of the DNN. Adapting the Army's DEBRIEF methodology, [17] adapted the assessment method for a human to better assess the performance of artificial intelligence (AI) (Figure 28).

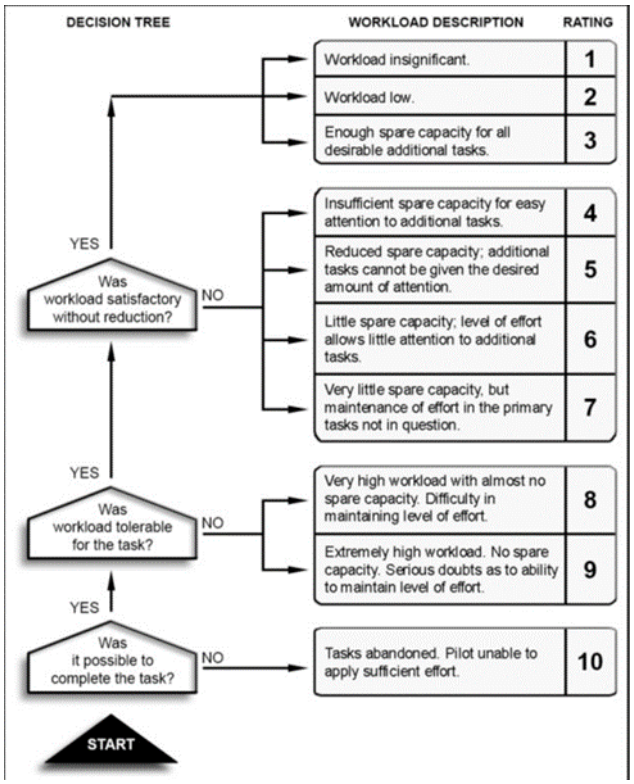

Figure 27. Revised Bedford Workload Scale. Source [28].

The first two steps in the process would be completed prior to execution of the test and for this use case, would remain constant. In the event that the DNN commits one of the six errors described in Figure 23, the test engineer could then complete steps three and four of the process. Now, given the speed of flight test, the ability to complete steps five and six may not be possible real-time if additional testing could be accomplished. In the event of a failure of the DNN and the decision is made to return to base (RTB), the test flight is far from over and the test engineers most likely will have to assist in monitoring the performance of the air vehicle during the RTB. Once safely on deck and video from

the flight may be reviewed, the remaining step (five, six, and seven) may be completed. While this use case has a relatively simple set of rules and objectives, this type of method will provide an increasing benefit as the complexity of the autonomous operation increases, such as cognitive electronic warfare or battlespace management, and this method should be adopted throughout the entire NAE.

Understanding that real-time labeling will not be possible to assess the metrics outlined in the simulator and laboratory testing segment, the ability exists through post-flight to download videos, label them, and then assess the performance of the DNN during specific parts of the flight. As new algorithms are developed to assist in labeling quickly, this process will become faster and will allow for more and more videos to be reviewed post-flight, but the understanding needs to be had that if manual labeling were to be relied on, a penalty in time would be added to the overall schedule of the test program. Now, if the domain changes, such as the performance in question occurred during a difficult environmental domain of the ODD and the following day was forecasted to be a more favorable environmental domain, the decision could be made to continue testing in support of the receiver certification program while a concurrent evaluation of the previous day's performance was conducted, but that type of concurrent evaluation would be contingent on available manpower.

AAR Steps		AAR/AI Questions Answered	AAR/AI Empirical Context
1. Define the rules.		How are we going to do this evaluation? What are the details regarding the situation?	We established the rules of evaluation and the domain (see the supplemental materials).
2. Explain the agent's objectives.		What is the AI's objective or objectives for this situation?	We explained the AI's objectives for the situation (see the supplemental materials).
AAR/AI Inner Loop	3. Review what was supposed to happen.	What did the evaluator intend to happen?	We asked, *"What do you think should happen in the next round(s)?"*
	4. Identify what happened.	What actually happened?	The participant watched the required number of rounds. Then, we asked, *"Could you briefly explain what actually happened in these past rounds?"*
	5. Examine why it happened.	Why did things happen the way they did?	We asked, *"Why do you think the rounds happened the way they did?"* Next, the participant summarized anything good, bad, or interesting on an index card. Last, we provided the participant the agent's explanation.
	6. Formalize learning (end inner loop).	Would the evaluator allow the AI to make these decisions on their behalf? What changes would they make in the decisions made by the AI to improve it?	We asked three questions: *"Would you allow the AI to make these decisions on your behalf?" "What changes would you make in the decisions made by the AI to improve it?" "Would you allow the Friendly AI to make this category of decisions on your behalf?"*
7. Formalize learning.		What went well, what did not go well, and what could be done differently next time?	The participant completed a post-task questionnaire (see the supplemental materials).

Figure 28. After-Action Review for AI. Source [17].

The final qualitative evaluation concept being covered in this research will be the mirroring of pilot training as a means to qualitatively assess the performance of a UAS. Traditionally, operational test plans consist of MOE and MOS, but with autonomous operation, consideration should be given to aligning MOE and MOS to traditional pilot qualifications [15]. For example, if the fleet replacement squadrons (FRS) carrier qualification syllabus called for 10-day carrier arrested landings and 4-night carrier arrest landings with a boarding rate of 80 percent, those metrics could be used as a means to operationally evaluate the performance of a UAS tasked with the same mission. Given the small scope of the DNN assurance test program though, this would be analogous to defining a metric for how many times an instructor should expect a student to correctly

identify key objects in the domain (100 percent ideally). Now, under the larger receiver certification program, a metric like this may be adopted mirroring the FRS's receiver qualification syllabus, but in terms of object detection, this type of comparison shows minimal value.

A key concept to understand is the risk that will need to be assigned to these types of capabilities (neural networks). A CBTE program could complete the entire receiver certification program, provide artifacts in support of assurance of the DNN, and a PFC could be issued to the UAS to allow it to conduct this mission within a specified envelope driven by the strengths and weaknesses discovered during the evaluation. On day one of the fleet using this UAS, for reasons unknown to the human monitoring the performance of the UAS in real-time, the system may fail to correctly identify the basket within the cleared envelope and be required to discontinue the mission. This echoes the concepts previously covered that a DNN has a near-infinite input space and even given the robust program outlined above, an edge case may be discovered in real-time that was not found during testing. In relation to the use of traditional pilot qualifications, a risk assessment is made by the instructing aircrew of the student in the given domain the instructor presented to the student. In the case of air-to-air refueling, the student pilot may satisfactorily complete the syllabus at the FRS and on the first day of deployment, fail to tank successfully due to an edge case not covered in training or evaluation. This type of risk assessment, while common in the pilot training program, would be new to an organization such as NAVAIR and therefore, a key consideration for the test needs to be providing sufficient artifacts to support this type of risk assessment.

So lastly, we have summarized the ways in which we might capture performance assessments of a UAS, via some standardized qualitative and quantitative methods, and how one might capture those activities in an after-action review. Let's summarize basics of the processes conferred as a baseline for discussions on broaching the subject of autonomous operations for unmanned vehicles, specifically in the case of air-to-air refueling.

THIS PAGE INTENTIONALLY LEFT BLANK

VII. CONCLUSION

To summarize, the complexity of modern warfare has rapidly outmatched the capacity of a human brain to accomplish the required tasks of a defined mission set. Task-shedding mundane tasks would prove immensely beneficial, freeing the warfighter to solve more complex issues; however, most tasks which a human might find menial, and shed-worthy, prove vastly abstract for a computer to solve. Advances in Deep Neural Network technology have demonstrated extensive applications as of recent. As DNNs become more capable of accomplishing increasingly complex tasks and the processors to run those neural nets continue to decrease in size, incorporation of DNN technology into legacy and next generation aerial Department of Defense platforms has become eminently useful and advantageous. The assimilation of DNN-based systems using traditional testing methods and frameworks to produce artifacts in support of platform certification within Naval, however proves prohibitive from a cost and time perspective, is not factored for agile development, and would provide an incomplete understanding of the capabilities and limitations of a neural network.

In order to begin the test and evaluation of unmanned and autonomous systems, a valid framework must first be established and proven. This framework, in the truest sense of the word, must set the formal—and hopefully lean—methods by which one proves a system's validity throughout the entirety of airworthiness process. This process must be compiled in such a way as to ensure that agility and responsiveness to the warfighter are at the forefront, with a dedication to delivering capabilities to the fleet vice becoming mired in processes. This is the only way we will adapt and overcome the increasing threat of great power conflict.

This research provides stakeholders within the naval aviation enterprise (NAE) and idea of the current tools, methodologies, and frameworks that exist or that are being developed, which, if adapted, could provide additional learning assurance for systems using machine learning to achieve a task currently reserved for a human operator. The primary purpose of the paper has succeeded and will be verified by actual tests within the coming years. This will be accomplished by examining the use case, specifically the

autonomous air-to-air refueling (A3R) of an unmanned aerial system (UAS) with a manned platform, and by relating how new capabilities in the field of assurance could be applied to the test planning, execution, and analysis process. The secondary objectives of capturing the processes, which will be a product of the actual testing and the research that goes into it, will make the same procedures extensible to future mission-sets suitable for autonomous systems. As the initial test and evaluation of this particular use-case proceeds, more will be learned, and, when captured, we will build the foundations for future test and evaluation.

The methods on how to test Unmanned Aerial System will continue to expand and become more complicated as the systems become more complex. As these systems become more complex, which they undoubtedly will as we strive to maintain parity with our adversaries, NAVAIR will at least have some tools in the proverbial toolbelt with which to succeed and build upon.

Future work areas involve the discovery of additional mission-sets that might be accomplished by Deep Neural Nets, while also meeting the needs of the Navy and matching National Security Strategy and National Defense Strategy goals. Mission might include, for example, the evaluation of autonomous platforms for Command, Control, Communications, Computers, and Intelligence, Surveillance, and Reconnaissance (C4ISR) roles while pushing advancements in resilient and agile logistics at a smaller scale. The possibilities are ultimately endless and merely require some dedicated work.

LIST OF REFERENCES

[1] J. K. Parry, "Framework Adaptations for Additional Assurance of a Deep Neural Network Within Naval Test and Evaluation," Ph.D. dissertation, School of Aviation and Trans. Tech., Purdue Univ., West Lafayette, Indiana, forthcoming.

[2] *Airworthiness and CYBERSAFE Process Manual*, M-13034.1, Department of the Navy, NAVAIR, Dec 2021.

[3] National Security Commission on Artificial Intelligence, "Final Report," Arlington, VA, USA, May 2021 [Online]. Available: https://www.nscai.gov/2021-final-report/

[4] White House, "Interim National Security Strategic Guidance," Washington, DC, USA, March 2021 [Online]. Available: https://www.whitehouse. gov/wp-content/uploads/2021/03/NSC-1v2.pdf

[5] Costello, D. H., and Adams, R., "A framework for airworthiness certification of autonomous systems within naval aviation," Jun 2021.

[6] Naval Air Systems Command, "Test planning manual version 2.0," Patuxent River, October 2018 [Online]. Available: https://directives.navair.navy.mil

[7] Purton, L., and Kourousis, K., "Military airworthiness management frameworks: a critical review," *Procedia Engineering*, vol. 80, Elsevier Ltd, 2014, pp. 545–564, September 2014.

[8] Public Affairs, "*Fueled In Flight: X-47B First To Complete Autonomous Aerial Refueling: NAVAIR*," Apr 2015 [Online]. Available: https://www.navair.navy.mil/node/22191

[9] *The Defense Acquisition System*, DOD Directive 5000.01, Department of Defense, 2020.

[10] *Test and Evaluation*, DOD Instruction 5000.89, Department of Defense 2020,

[11] Defense Acquisition University. *DOD Test and Evaluation Organizations*. Accessed: 2022 [Online]. Available: https://www.dau.edu/tools/Lists/DAUTools/Attachments/148/Test%20and%20Evaluation%20Management%20Guide,%20December%202012,%206th%20Edition%20-v1.pdf

[12] Defense Acquisition University, "*Test and evaluation management guide*," Fort Belvoir, VA, Sixth Edition, 2012.

[13] Defense Acquisition University, “Defense Acquisition Guidebook,” 2020 [Online]. Available: https://www.dau.edu/tools/ t/Defense-Acquisition-Guidebook.

[14] Bjorkman, E. A., Sarkani, S., and Mazzuchi, T. A., “Test and evaluation resource allocation using uncertainty reduction,” *IEEE Transactions on Engineering Management*, Vol. 60, No. 3, pp. 541–551. 2013

[15] *Test and Evaluation of Autonomy for Air Platforms*, Livermore, R. A., and Leonard, A. W., 412th Test Wing, Edwards Air Force Base, California, 2020.

[16] Lednicky, E. J., and Silvestrini, R. T., “Quantifying gains using the capabilities-based test and evaluation method,” *Quality and Reliability Engineering International*, Vol. 29, No. 1, pp. 139–156, 2013.

[17] Dodge, J., Khanna, R., Irvine, J., Lam, K.-h., Mai, T., Lin, Z., Kiddle, N., Newman, E., Anderson, A., Raja, S., Matthews, C., Perdriau, C., Burnett, M., and Fern, A., “After-action review for ai (aar/ai),” *ACM Trans. Interact. Intell. Syst.*, Vol. 11, No. 3–4, 2021 [Online]. Available: https://doi.org/10.1145/3453173.

[18] United States Air Force, *KC-46A pegasus*, January 2019 [Online]. Available: https://www.af.mil/About-Us/Fact-Sheets/Display/Article/104537/kc-46a-pegasus/.

[19] *Automated Air-To-Air Refuelling*, ATP-3.3.4.10, North Atlantic Treaty Organization, 2021

[20] “OpenCV AI Kit: OAK-D,” 2021 [Online]. Available: https://store.opencv.ai/products/oak-d.

[21] Unknown, *FA 18 Carrier Flying VFA 154 Black Knight Cruise 2021 4K*, Accessed: 2021 [Online Video]. Available: https://www.youtube.com/watch?v=kO8QSTmdwCc

[22] *Air-to-Air Refuelling*, ATP 3.3.4.2.(D), North Atlantic Treaty Organization, 2021

[23] Bharadwaj, R., “Assuring autonomy,” *Disruptive Technologies in Information Sciences IV*, Vol. 11419, p. 114190G, International Society for Optics and Photonics, 2020

[24] Fisher, M., Mascardi, V., Rozier, K. Y., Schlingloff, B.-H., Winikof, M., and Yorke-Smith, N., “Towards a framework for certification of reliable autonomous systems,” *Autonomous Agents and Multi-Agent Systems*, Vol. 35, No. 1, 2021.

[25] Liu, C., Arnon, T., Lazarus, C., Strong, C., Barrett, C., and Kochenderfer, M. J., “Algorithms for verifying deep neural networks,” 2020 [Online]. Available: https://github.com/sisl/NeuralVerification.jl

[26] Ratto, C., Pekala, M., Fendley, N., Drenkow, N., Karra, K., Ashcraft, C., Costello, C., Burlina, P., Wang, I.-J., and Wolmetz, M., "Adversarial Machine Learning and the Future Hybrid Battlespace," The Johns Hopkins University Applied Physics Laboratory.

[27] Thorn, E., Kimmel, S., and Chaka, M., "A framework for automated driving system testable cases and scenarios," NHTSA, 2018.

[28] Nickell, C., "Aerial Refueling Receiver Certification Test Methodology," 2021.

[29] MVRsimulation, "Virtual reality scene generator (Vrsg) Overview," 2021 [Online] Available: https://www.mvrsimulation.com/products/vrsg-overview.html

[30] Abu Alhaija, H., Mustikovela, S. K., Mescheder, L., Geiger, A., and Rother, C., "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *International Journal of Computer Vision*, Vol. 126, No. 9, pp. 961–972, 2018.

[31] Wang, X., Dong, X., Kong, X., Li, J., and Zhang, B., "Drogue detection for autonomous aerial refueling based on convolutional neural networks," *Chinese Journal of Aeronautics*, Vol. 30, No. 1, pp. 380–390, 2017 [Online]. Available: https://doi.org/10.1016/j.cja.2016.12.022

[32] Delseny, H., Gabreau, C., Gauffriau, A., Beaudouin, B., Ponsolle, L., Alecu, L., Bonnin, H., Beltran, B., Duchel, D., Ginestet, J.-B. et al., "White paper machine learning in certified systems," 2021 [Online]. Available: arXiv:2103.10529

[33] Gopinath, D., Katz, G., Păsăreanu, C. S., and Barrett, C., "Deepsafe: A data-driven approach for assessing robustness of neural networks*," International Symposium on Automated Technology for Verification and Analysis*, pp. 3–19, 2018 [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-01090-4_1#citeas

[34] Katz, G., Barrett, C., Dill, D. L., Julian, K., and Kochenderfer, M. J., "Reluplex: An efficient SMT solver for verifying deep neural networks," *International conference on computer aided verification*, 2017 [Online]. Available: arXiv:1702.0113

[35] Goldsborough, P., "A tour of tensorflow," 2016 [Online]. Available: arXiv:1610.01178

[36] Ayers, E. W., Eiras, F., Hawasly, M., and Whiteside, I., "PaRoT: a practical framework for robust deep neural network training," *NASA Formal Methods Symposium*, 2020 [Online]. Available: arXiv:2001.02152

[37] Bak, S., Liu, C., and Johnson, T., “The second international verification of neural networks competition (vnn-comp 2021): summary and results,” 2021 [Online]. Available: arXiv:2109.00498

[38] NASA Formal Methods 13th International Symposium, NFM 2021, Virtual Event, May 24–28, 2021, Proceedings, 1st ed., Programming and Software Engineering; 12673, Springer International Publishing, Cham, 2021.

[39] NASA Formal Methods 12th International Symposium, NFM 2020, Moffett Field, CA, USA, May 11–15, 2020, Proceedings, 1st ed., Programming and Software Engineering; 12229, Springer International Publishing, Cham, 2020.

[40] NASA Formal Methods 11th International Symposium, NFM 2019, Houston, TX, USA, May 7–9, 2019, Proceedings, 1st ed., Programming and Software Engineering; 11460, Springer International Publishing, Cham, 2019.

[41] NASA Formal Methods 10th International Symposium, NFM 2018, Newport News, VA, USA, April 17–19, 2018, Proceedings, 1st ed., Programming and Software Engineering; 10811, Springer International Publishing, Cham, 2018.

[42] NASA Formal Methods 9th International Symposium, NFM 2017, Moffett Field, CA, USA, May 16–18, 2017, Proceedings, 1st ed., Programming and Software Engineering; 10227, Springer International Publishing, Cham, 2017.

[43] Verdhan, V., *Computer Vision Using Deep Learning*. Apress, 2021 [Online]. Available: https://doi.org/10.1007/978-1-4842-6616-8

[44] Wenkel, S., Alhazmi, K., Liiv, T., Alrshoud, S., and Simon, M., “Confidence score: The forgotten dimension of object detection performance evaluation,” *Sensors (Basel, Switzerland)*, Vol. 21, pp. 4350–4371, 2021 [Online]. Available: https://doi.org/10.3390/s21134350

[45] Ma, Y., Zhao, R., Liu, E., Zhang, Z., and Yan, K., “A novel autonomous aerial refueling drogue detection and pose estimation method based on monocular vision,” *Measurement: Journal of the International Measurement Confederation*, Vol. 136, pp. 132–142, 2021 [Online]. Available: https://doi.org/10.1016/j.measurement.2018.12.060

[46] Li, W., “Analysis of object detection performance based on faster R-CNN,” *Journal of Physics: Conference Series*, Vol. 1827, IOP Publishing Ltd, 2021 [Online]. Available: https://doi.org/10.1088/1742-6596/1827/1/012085

[47] Bolya, D., Foley, S., Hays, J., and Hoffman, J., “Tide: A general toolbox for identifying object detection errors,” *European Conference on Computer Vision*, 2020 [Online]. Available: arXiv:2008.08115

[48] Bownes, Vincent J., "Using motion capture and augmented reality to test aar with boom occlusion" (2021). Theses and Dissertations, AFIT, 2021 [Online]. Available: https://scholar.afit.edu/etd/4993

[49] Hansen, J., Romrell, G., Nabaa, N., Andersen, R., Myers, L., and McCormick, J., "DARPA autonomous airborne refueling demonstration program with initial results," *Proceedings of the 19th International Technical Meeting of the Satellite Division of The Institute of Navigation* (ION GNSS 2006), pp. 674–685, 2006 [Online]. Available: https://www.ion.org/publications/abstract.cfm?articleID=6897

[50] Insinna, V., "The air force's first skyborg autonomous drone prototype made its first flight," *DefenseNews*, 2021 [Online]. Available: https://www.defensenews.com/air/2021/05/05/the-air-forces-first-skyborg-autonomous-drone-prototype-made-its-first-flight/

[51] Tirpak, J. A., "Contracts advance skyborg toward becoming program of record on time," Air Force Magazine, 2021 [Online]. Available: https://www.airforcemag.com/skyborg-air-force-contracts-advance/

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California