

# DS IAC JOURNAL

Aligning the Navy's  
Fleet Experimentation With  
Capability Development  
**PAGE 4**

Screening Low-Flammability  
Polymers Through  
Chemistry-Aware ML  
**PAGE 14**

Multiagent Federated  
Learning, Interoperability,  
and Virtual-Physical  
Co-simulation  
**PAGE 36**

Seminal Technique for  
Characterizing GPS  
Navigation  
Performance  
**PAGE 46**

**PAGE 25**  
**DATA-OPTIMIZED HUMAN-  
MACHINE TEAMING WITH  
ROBOTIC WINGMEN**



**Editor-in-Chief:**

Aaron Hodges

**Sr. Technical Editor:**

Maria Brady

**Graphic Designers:**

Melissa Gestido, Katie Ogorzalek

The DSIAC Journal is a publication of the Defense Systems Information Analysis Center (DSIAC). DSIAC is a DoD Information Analysis Center (IAC) sponsored by the Defense Technical Information Center (DTIC) with policy oversight provided by the Office of the Under Secretary of Defense (OUSD) for Research and Engineering (R&E). DSIAC is operated by the SURVICE Engineering Company.

Copyright © 2025 by the SURVICE Engineering Company.

This journal was developed by SURVICE under DSIAC contract FA8075-21-D-0001. The Government has unlimited free use of and access to this publication and its contents, in both print and electronic versions. Subject to the rights of the Government, this document (print and electronic versions) and the contents contained within it are protected by U.S. copyright law and may not be copied, automated, resold, or redistributed to multiple users without the written permission of DSIAC. If automation of the technical content for other than personal use, or for multiple simultaneous user access to the journal, is desired, please contact DSIAC at 443.360.4600 for written approval.

Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or DSIAC.

The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or DSIAC and shall not be used for advertising or product endorsement purposes.

ISSN 2471-3392 (Print) // ISSN 2471-3406 (Online)

**Distribution Statement A:**

Approved for public release; distribution is unlimited.

**On the Cover:**

Digital Art Rendering (Source: Canva and Adobe Stock).



## ABOUT DSIAC

### Who We Are

A DoD Information Analysis Center comprised of scientists, engineers, researchers, analysts, and information specialists.

### What We Do

Generate, collect, research, analyze, synthesize, and disseminate scientific and technical information (STI) to DoD and federal government users and industry contractors.

### Why Our Services

To eliminate redundancy, foster collaboration, and stimulate innovation.

## DSIAC SERVICES



### Subject Matter Expert (SME) Connections

Access to a network of experts with expertise across our technical focus areas.



### Technical Inquiries (TIs)

Up to 4 hours of FREE research using vast DoD information resources and our extensive network of SMEs.



### Specialized Task Orders

Research and analysis services to solve our customer's toughest scientific and technical problems.



### Webinars & Events

Our webinars feature a technical presentation from a SME in one of our focus areas. We also offer key technical conferences and forums for the science and technology community.



### STI Collection

Our knowledge management team collects and uploads all pertinent STI into DTIC's Research & Engineering Gateway.



### Information Research Products

The Defense Systems Digest, state-of-the-art reports, journals, TI response reports, and more available on our website.

## CONTACT DSIAC

### IAC Program Management Office

8725 John J. Kingman Road  
Fort Belvoir, VA 22060  
**Office:** 571.448.9753

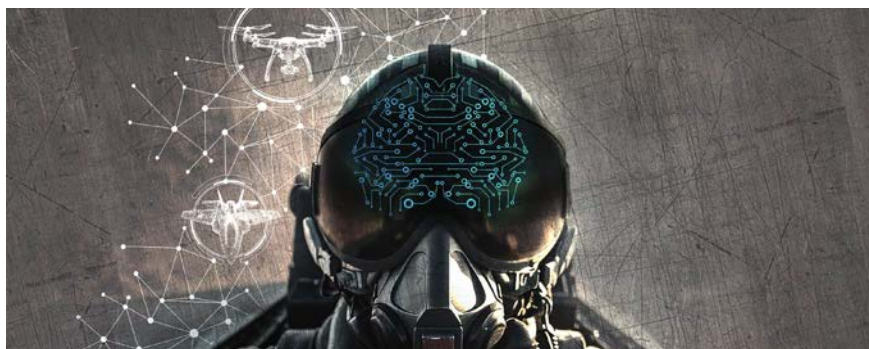
### DSIAC Headquarters

4695 Millennium Drive  
Belcamp, MD 21017-1505  
**Office:** 443.360.4600  
**Fax:** 410.272.6763  
**Email:** [contact@dsiac.org](mailto:contact@dsiac.org)

### DSIAC Technical Project Lead

Taylor Knight  
4695 Millennium Drive  
Belcamp, MD 21017-1505  
**Office:** 443.360.4600





## DATA-OPTIMIZED HUMAN-MACHINE TEAMING WITH ROBOTIC WINGMEN

By Christina Hayhurst, Christine Covas-Smith, and Patricia Harris

This article will recommend a collaborative combat aircraft (CCA) data readiness tool that quickly informs commanders of their level of risk assumption based on data criteria rankings and using the tool's application

to cognitive electronic warfare (CEW) CCAs as a vignette. It will also suggest the personnel, policy, and technology investments to optimize each data readiness criterion.

## IN THIS ISSUE

### 04 Aligning the Navy's Fleet Experimentation With Capability Development

By Jonathan Haase

### 14 Screening Low-Flammability Polymers Through Chemistry-Aware ML

By George M. Nishibuchi, Suhas Chelian, Wyler Zahm, Srinivasan, and Richard E. Lyon

### 36 Multiagent Federated Learning, Interoperability, and Virtual-Physical Co-simulation

By Nirmalya Roy, Jade Freeman, Mark Dennison, Theron Trout, and Timothy Gregory

### 46 Seminal Technique for Characterizing GPS Navigation Performance

By Nathan A. Ruprecht, Elisa N. Carrillo, and Loren E. Myers



## Investigating Low Global-Warming-Potential Refrigerants for Military Transport Applications

By Claire E. O'Malley, Changkuan Liang, Enrique A. Velazquez, Robert E. Ferguson, Listier A. Otieno, Steven F. Son, and Davide Ziviani | Photo Source: Canva

This study investigates minimally flammable alternatives to R-134a and R-1234yf by using a holistic approach that encompasses both thermodynamic aspects and flammability characteristics.



**AVAILABLE ONLY ONLINE**

<https://buff.ly/bDrlj7h>



# Aligning the Navy's Fleet Experimentation With Capability **DEVELOPMENT**



**BY JONATHAN HAASE**

(PHOTO SOURCE: ADOBE  
STOCK, CANVA, AND 123RF.COM)



# INTRODUCTION

**A**midst the crashing surf of a remote coastline, two U.S. Navy sailors clad in black wetsuits and protective helmets carry the MK18 Mod 1 Swordfish unmanned underwater vehicle (UUV) toward the open ocean (Figure 1). With minimal support equipment, they navigate the surf zone, demonstrating the expeditionary capability of deploying UUVs virtually anywhere in the world. Nearby, another sailor hunches over a rugged laptop, programming the UUV's mission parameters in the open air—a testament to the simplicity and flexibility of operating these advanced systems in austere environments.

This scene unfolded during UNITAS LXV, the world's longest-running multinational maritime exercise (Figure 2), highlighting the tactical proficiency of America's warfighting Navy. The deployment of the MK18 Mod 1 Swordfish by Explosive Ordnance Disposal Mobile Unit 2 not only showcased cutting-edge technology but also emphasized the importance of interoperability with allies and partners such as Canada, France, and Germany. Such exercises promote peace, stability, and prosperity by fostering collaboration and enhancing collective maritime capabilities (Figure 3).

The MK18 Mod 1 Swordfish's successful deployment underscores the



**Figure 1.** UUV Operations (Source: DVIDS [1]).



**Figure 2.** UUV Land Operation at UNITAS LXV (Source: DVIDS [1]).

transformative potential of UUVs in modern naval operations. Their ability to be rapidly deployed and operated with minimal logistical support makes them ideal assets for a wide range of missions, from mine countermeasures to intelligence gathering. This agility is critical in today's fast-paced, evolving threat environment.

This article explores how leveraging contractor-owned, contractor-operated (COCO) services, in combination with fleet experimentation like that demonstrated during UNITAS LXV, can accelerate Programs of Record (PoRs) and better align capabilities with fleet needs. By tightening the feedback loop between sailors and program managers, the Navy ensures



**Figure 3.** UUV Familiarization Training (Source: DVIDS [1]).

an active role in advancing operational requirements. This approach fosters innovation, enhances operational effectiveness, and positions the Navy to meet current and future challenges with agility and precision.

## Evolving Needs in Capability Development

The U.S. Department of the Navy has a longstanding tradition of innovation and excellence in fielding critical systems that safeguard the nation's interests. Its traditional acquisition processes have been instrumental in delivering robust and reliable capabilities that have served well over the years. These processes ensure thorough evaluation, accountability, and adherence to high standards, which are essential for mission success and the safety of Navy personnel.

However, the rapidly changing technological landscape and

emergence of new threats presents an opportunity to enhance existing processes. As technology evolves at an unprecedented pace, the methods for developing and deploying new capabilities must also evolve. By building upon the strong foundation of the current acquisition framework, more agile and responsive approaches can be integrated to effectively meet today's challenges.

One area that can be advanced is optimizing development timelines. While thoroughness is crucial, exploring ways to streamline certain aspects of the acquisition process can help bring cutting-edge technologies to sailors more quickly. This does not mean compromising on quality or safety but finding efficiencies that allow faster integration of new solutions without sacrificing rigor.

Additionally, introducing greater flexibility into the processes can

enhance the ability to adapt to emerging technologies and evolving operational requirements. By complementing structured acquisition methods with more dynamic strategies, response to new opportunities can be swift and capabilities remaining at the forefront of innovation can be ensured.

Strengthening the alignment between program development and fleet needs is another avenue for positive growth. By fostering closer collaboration and open communication between program managers, industry partners, and the sailors who will ultimately employ these systems, the capabilities developed to address the most pressing operational challenges can also be ensured.

In embracing these enhancements, maintaining the excellence of these acquisition processes while making them more agile and responsive is the goal. This balanced approach allows continued deliverance of superior capabilities that empower sailors and keep them ahead in an ever-changing global landscape.

## The Need for Agility in Modern Warfare

Modern warfare is evolving at an unprecedented pace, driven by rapid advancements in technology and shifting battlefield dynamics. Unmanned systems and artificial intelligence (AI) are at the forefront of this transformation, redefining

“

***Modern warfare is evolving at an unprecedented pace, driven by rapid advancements in technology and shifting battlefield dynamics.***

how conflicts are conducted and highlighting the critical need for agility in capability development.

As drone technology evolves, so do the strategies employed on the battlefield. Unmanned systems offer significant advantages, including reduced risk to personnel, enhanced surveillance capabilities, and the ability to respond rapidly to emerging threats. The statement “the future of warfare is here, and it’s unmanned” resonates strongly in this context, emphasizing the immediate impact of these technologies on modern conflicts.

This evolving landscape underscores the importance for the Navy to adopt more agile approaches to capability development. By embracing flexibility and innovation, the Navy can ensure that it remains ahead of emerging threats and continues to provide sailors with the most advanced tools available. Agile development processes allow quicker integration of new technologies like unmanned systems and AI, ensuring operational readiness and maintaining technological superiority.

Incorporating lessons from real-world examples, such as the use of drones in Ukraine, highlights the necessity of adapting our acquisition strategies. By fostering agility, the Navy can better align its capabilities with the fast-paced nature of modern warfare, ensuring that it is prepared to meet current and future challenges with confidence and precision.

## THE ROLE OF FLEET EXPERIMENTATION

Testing, learning, and improving are crucial for advancing naval capabilities, thus allowing the following:

- **Real-World Testing:** Operational environments provide valuable data on system performance, usability, and reliability.
- **Rapid Feedback Loops:** Direct input from sailors enables program managers to make informed decisions quickly.
- **Iterative Improvement:** Continuous testing and learning lead to refined systems that better meet operational needs.

By integrating experimentation into the capability development process, the Navy can provide the following benefits:

- **Accelerate Innovation:** Quickly identify and adopt emerging technologies.

- **Enhance Relevance:** Ensure that capabilities are directly aligned with fleet needs.
- **Reduce Risk:** Validate concepts and technologies before full-scale acquisition.

## LEVERAGING COCO SERVICES

COCO services represent a collaborative model where contractors own and operate equipment or provide services, delivering capabilities directly to the Navy without the need for the government to procure the assets outright. This approach offers the following key advantages:

- **Flexibility:** Contractors can rapidly design, experiment, and test ideas independently, allowing swift adaptation to emerging technologies and operational needs.
- **Cost Efficiency:** With reduced government oversight, timelines accelerate and administrative costs decrease, leading to more efficient use of resources.
- **Innovation:** By encouraging industry partners to bring forward cutting-edge solutions, the Navy benefits from the latest advancements without bearing the full burden of development risks.

COCO services provide the following advantages:



- **Speed of Implementation:** Vendors can iterate quickly without being hindered by lengthy government approval processes. This agility enables the Navy to field new capabilities more rapidly in response to evolving threats.
- **Reduced Costs:** Lower requirements for government-funded testing and oversight translate into significant cost savings. Funds can be redirected toward other critical areas without compromising capability development.
- **Enhanced Collaboration:** Strong partnerships between the Navy and industry are fostered, leveraging external expertise and promoting a shared commitment to advancing naval capabilities.

## Case Study: Unmanned Underwater Vehicles in BALTOPS 24 [2]

A recent example highlighting the effectiveness of COCO services occurred during BALTOPS 24, the world's longest-running multinational maritime exercise. The Royal Netherlands Navy deployed a yellow UUV from one of their ships to investigate underwater contacts as part of mine countermeasure training (Figure 4).

This deployment showcased the following important aspects:

- **Operational Value:** The UUV operated effectively in a real-world

“

*A recent example highlighting the effectiveness of COCO services occurred during BALTOPS 24, the world's longest-running multinational maritime exercise.*

training environment, demonstrating its capability to enhance mine detection and clearance operations.

- **Rapid Deployment:** The ability to launch the UUV swiftly during an international exercise underscored the flexibility and speed that COCO services could provide.
- **Industry and International Collaboration:** The exercise highlighted how contractors, allied navies, and partners were willing and able to collaborate, bringing

forward innovative technologies that enhanced collective maritime security.

By integrating COCO services into such exercises, the Navy and its partners can evaluate new technologies in operational settings without the need for immediate procurement. This approach allows assessing capabilities, identifying any limitations, and gathering valuable feedback from operators.

## The Triton UUV and Combined Task Force (CTF) 59 Operations

Another compelling example is the Triton UUV's operations in the Arabian Gulf. Despite the government not purchasing Tritons specifically for these activities, Ocean Aero demonstrated the UUV's capabilities through fleet exercises and demonstrations coordinated



**Figure 4.** Sea Scan Underwater Drone (Source: DVIDS [2]).



by CTF 59 [3]. This collaboration showcased the following:

- **Proven Endurance:** The Triton UUV operated for extended periods, validating that long-duration unmanned systems could function reliably in challenging maritime environments.
- **Unified Control Systems:** CTF 59 successfully integrated multiple unmanned platforms, controlling various UUVs from a single interface. This capability enhanced operational efficiency and situational awareness.
- **Industry Engagement:** Contractors proactively participated to prove the potential value of their technologies, reinforcing the benefits of COCO services in advancing naval capabilities.

## INTEGRATING COCO SERVICES WITH PoRs

It is a connection most glossed over, but when COCO services are viewed as a path to rapid PoR advances, new and useful opportunities are presented. Understanding the flexibility of COCO services is key to using these services to rapidly advance PoRs.

### Bridging the Gap Between Experimentation and Acquisition

To fully harness the benefits of COCO services, it is crucial to understand

how they can transition into formal PoRs. The REMUS 300 UUV serves as an excellent example of a commercial system that bridges this gap. Developed by Huntington Ingalls Industries (HII), the REMUS 300 is being used to establish a PoR and could further benefit from COCO services to implement incremental improvements [4].

Key considerations in this transition include the following:

- **Data Rights:** Ensuring that the Navy has access to essential data generated during COCO operations is vital. This data supports evaluation, informs future development, and ensures that any enhancements align with Navy requirements.
- **Test and Evaluation Requirements:** Establishing clear protocols for testing and validating performance is necessary to ensure that systems meet operational standards and are suitable for deployment.
- **Quality Evidence Requirements:** Gathering objective and quantifiable data provides the evidence needed to support acquisition decisions and justifies the integration of new capabilities into PoRs.

### Enhancing the Acquisition Process

Integrating COCO services with PoRs offers several advantages that enhance the traditional acquisition process:

- **Informing PoRs:** COCO services provide real-world operational data and direct fleet feedback. For instance, using the REMUS 300 in COCO arrangements allows sailors to operate the system in various missions such as mine countermeasures, data collection, and search and rescue. Their experiences help shape requirements and specifications for PoRs.
- **Reducing Risks:** Early identification and mitigation of potential issues are possible when systems are tested extensively in operational environments. Incremental improvements made through COCO services can be evaluated before formal incorporation into PoRs, reducing development risks.
- **Accelerating Timelines:** By utilizing commercial systems, which are already operationally viable, the Navy can shorten the path from concept to deployment. COCO services facilitate rapid iteration and integration of new technologies without the delays often associated with traditional procurement processes.

### Ensuring Alignment With Fleet Needs

Direct involvement of the fleet in experimentation ensures that the following capabilities developed are precisely what sailors require:

- **Operational Relevance:** Systems are tested and validated by end

users in real-world scenarios. For example, during mine countermeasure operations, sailors can assess the UUV's effectiveness and suggest enhancements based on firsthand experience.

- **Responsive Development:** Feedback from sailors operating COCO systems can lead to quick adjustments and incremental improvements. HII's recent unveiling of the REMUS 130, built on the same platform as the REMUS 300, exemplifies how industry can respond to operational feedback by offering vehicles with reduced cost and risk while maintaining high capability.
- **Active Navy Role:** By engaging directly with contractors through COCO services, the Navy maintains ownership of capability development priorities. This active role ensures that advancements align with strategic objectives and operational requirements.

### Case Study: The REMUS 300 and PoR Establishment

The REMUS 300's journey from a commercial UUV to its role in establishing a PoR illustrates the effective integration of COCO services as follows:

- **Commercial Foundation:** The REMUS 300, a small UUV measuring between 6 to 12 ft long and 7 1/2 inches in diameter, is designed for versatility in missions like data collection, offshore

exploration, search and rescue, and mine countermeasures.

- **Incremental Improvements:** HII's introduction of the REMUS 130, based on the proven REMUS 300 platform, demonstrates how incremental enhancements can be developed in response to customer needs.
- **Establishing the Lionfish UUV:** The REMUS 300 is also the commercial system being utilized by HII to develop and manufacture the Navy's new Lionfish UUV. This progression from a commercial product to a tailored military asset underscores the potential of leveraging COCO services for capability development.

## BENEFITS OF A COMBINED APPROACH

Integrating COCO services with traditional PoRs offers a multitude of benefits that enhances the Navy's capability development process. This combined approach leverages the strengths of both models to deliver advanced capabilities more efficiently and effectively.

### Increased Agility and Responsiveness

- **Adapts to Emerging Threats:** By utilizing COCO services, the Navy can rapidly develop and deploy new technologies to address evolving

“

*Integrating COCO services with traditional PoRs offers a multitude of benefits that enhances the Navy's capability development process.*

challenges. For example, the deployment of unmanned systems like the MK18 Mod 1 Swordfish during UNITAS LXV and the REMUS 300 UUV allows for swift integration of advanced capabilities to counter modern threats.

- **Facilitates Flexible Contracting:** COCO services enable the Navy to adopt diverse contracting strategies tailored to specific needs. This flexibility facilitates quicker procurement and deployment of essential technologies, ensuring that operational units have access to the tools they require without delay.

### Improved Alignment and Relevance

- **Directly Addresses Fleet Needs:** Involving sailors directly in the experimentation and development process ensures that capabilities are closely aligned with actual operational requirements. The hands-on use of UUVs like the REMUS 300 and the MK18 Mod 1 by fleet personnel provides



immediate feedback, allowing adjustments that meet the specific needs of the fleet.

- **Enhances Effectiveness:** Systems developed through this collaborative approach are more likely to perform as required in real-world scenarios. The practical experience gained from exercises like BALTOPS 24, where unmanned underwater vehicles were deployed for mine countermeasure training, demonstrates the operational effectiveness of these systems.

## Cost Savings and Efficiency

- **Reduces Development Costs:** Streamlined processes inherent in COCO services eliminate unnecessary expenses associated with traditional acquisition methods. By leveraging commercial off-the-shelf technologies and industry expertise, the Navy can reduce research and development costs significantly.
- **Efficiently Uses Resources:** Focusing funding on high-impact areas ensures that resources are utilized where they can make the most difference. The incremental improvements made to systems like the REMUS 300, leading to the development of the REMUS 130, exemplify efficient resource allocation that meets mission needs without excessive expenditure.

## Enhanced Collaboration and Innovation

- **Partners With Industry:** COCO services foster strong partnerships between the Navy and private sector companies. By leveraging advancements and expertise from industry leaders like HII, the Navy benefits from cutting-edge technologies and innovative solutions that might not be readily available through traditional procurement channels.
- **Fosters an Innovation Culture:** Encouraging creative problem-solving and adopting new technologies becomes a natural outcome of this combined approach. The agile development seen in the use of first-person-view drones by operators in Ukraine highlights how embracing new methodologies can lead to significant tactical advantages.

## ADDRESSING CHALLENGES AND MITIGATING RISKS

While the combined approach of integrating COCO services with PoRs offers substantial benefits, it is essential to recognize potential challenges and implement strategies to mitigate risks effectively. Such challenges include the following:

- **Misappropriation Concerns:** Ensuring the proper use of funds

and resources is crucial. Without adequate oversight, there is a risk of misaligned priorities or inefficient use of financial resources.

- **Unfunded Requirements:** Successful demonstrations and experiments may lead to identifying capabilities that lack allocated budgets for full-scale implementation, potentially causing gaps in operational readiness.
- **Testing Considerations:** Robust protocols are necessary to validate the effectiveness and reliability of new systems. Without standardized testing, there may be inconsistencies in performance or unforeseen issues during deployment.
- **Program Configuration Management:** Maintaining consistency with broader program objectives is vital. Changes or improvements made through COCO services must be carefully managed to ensure they align with the overall goals and do not disrupt existing configurations.

Strategies for mitigation are as follows:

- **Clear Guidelines and Oversight Mechanisms:** Implementing comprehensive frameworks for accountability ensures that all parties understand their responsibilities. Establishing clear contracts and expectations with industry partners helps maintain focus and proper use of resources.

- **Phased Funding Approaches:** Allocating resources based on milestones and performance allows better financial control. By tying funding to specific achievements, the Navy can ensure that investments lead to tangible results without overcommitting resources prematurely.
- **Collaboration With Program Managers:** Close coordination between COCO service providers and program managers ensures alignment with overall program goals. Regular communication and joint planning help integrate new capabilities smoothly into existing structures.
- **Standardized Testing Protocols:** Establishing consistent evaluation methods guarantees that all new systems meet required standards. Developing and adhering to rigorous testing procedures reduces the risk of deploying unproven technologies and enhances overall reliability.

## CONCLUSIONS

The landscape of modern warfare is rapidly evolving, driven by technological advancements and emerging threats that demand agility, innovation, and swift adaptation. The Navy stands at a pivotal juncture where embracing new methodologies can significantly enhance its operational effectiveness and maintain its strategic edge.

By integrating COCO services with traditional PoRs, the Navy can leverage the best of both worlds—harnessing industry innovation and flexibility while maintaining the rigorous standards and oversight that ensure mission success and safety of its personnel. This combined approach offers the following benefits:

- **Increased Agility and Responsiveness:** The ability to rapidly develop and deploy new technologies allows the Navy to address emerging challenges promptly. Flexible contracting strategies enable tailored solutions that meet specific operational needs without unnecessary delays.
- **Improved Alignment and Relevance:** Direct involvement of sailors in experimentation ensures that capabilities are developed with immediate operational input. This hands-on engagement enhances the likelihood that systems will perform as required in real-world scenarios, directly addressing the fleet's needs.
- **Cost Savings and Efficiency:** Streamlined processes inherent in COCO services reduce development costs and focus funding on high-impact areas. Efficient use of resources ensures that the Navy can invest in critical capabilities without excessive expenditure.
- **Enhanced Collaboration and Innovation:** Strong partnerships with industry leverage private sector advancements and expertise.

Fostering a culture of innovation encourages creative problem-solving and the adoption of cutting-edge technologies.

The examples discussed—from the deployment of the MK18 Mod 1 Swordfish during UNITAS LXV to the incremental improvements of the REMUS 300 UUV—illustrate the tangible benefits of this integrated approach. These cases highlight how the Navy can rapidly field advanced capabilities, improve operational readiness, and stay ahead of adversaries who are also leveraging technology to their advantage.

However, it is essential to recognize and address potential challenges associated with this approach. Implementing clear guidelines, robust oversight mechanisms, phased funding, and standardized testing protocols ensures that risks are mitigated effectively. Close collaboration between program managers, industry partners, and operational units maintains alignment with broader objectives and sustains the integrity of capability development efforts.

## THE PATH FORWARD

Embracing the integration of COCO services with PoRs represents a strategic imperative for the Navy. This approach not only accelerates capability development but also strengthens the Navy's active role in advancing operational requirements.



“

**Embracing the integration of COCO services with PoRs represents a strategic imperative for the Navy.**

By fostering agility, enhancing collaboration, and prioritizing innovation, the Navy positions itself to meet the demands of modern warfare confidently and effectively.

Moving forward, the Navy should continue to do the following:

- **Promote Agile Development**  
**Practices:** Encourage the adoption of flexible methodologies that allow rapid iteration and deployment of new technologies.
- **Strengthen Industry Partnerships:**  
Cultivate relationships with industry leaders to leverage expertise and stay abreast of technological advancements.
- **Enhance Fleet Engagement:**  
Involve sailors directly in the development and experimentation process to ensure capabilities align with operational realities.
- **Implement Robust Governance:**  
Establish frameworks that balance innovation with accountability, ensuring that resources are used efficiently and effectively.

By committing to these principles, the Navy can navigate the complexities of today’s security environment, maintaining technological superiority and operational excellence. The integration of COCO services with traditional acquisition processes is not just an opportunity but a necessity to ensure that the Navy remains agile, responsive, and prepared to face the challenges of the future. ■

## REFERENCES

- [1] Defense Visual Information Distribution Service (DVIDS). “Unmanned Underwater Vehicle Land Operations” Photo by Petty Officer 1st Class Hunter Harwell, U.S. Naval Forces Southern Command/U.S. 4th Fleet, <https://www.dvidshub.net/image/8628508/unmanned-underwater-vehicle-land-operations>, 2024.
- [2] DVIDS. “Sea Scan Underwater Drone Used for Training During BALTOPS 24” Courtesy photo by Royal Netherlands Navy Senior Chief Petty Officer Jan Eenling, U.S. Naval Forces Europe-Africa/U.S.

6th Fleet, <https://www.dvidshub.net/image/8469011/sea-scan-underwater-drone-used-training-during-baltops-24>, 2024.

[3] U.S. Department of Defense. “U.S., UAE Naval Forces Complete First-Ever Bilateral Unmanned Exercise.” U.S. Naval Forces Central Command Public Affairs, U.S., UAE Naval Forces Complete First-Ever Bilateral Unmanned Exercise > U.S. Naval Forces Central Command > Display, 20 February 2023.

[4] HII. “HII’s REMUS 300 Selected as U.S. Navy’s Next Generation Small UUV Program of Record!” <https://hii.com/news/hii-remus-300-selected-as-u-s-navys-next-generation-small-uuv-program-of-record/>, 20 March 2022.

## BIOGRAPHY

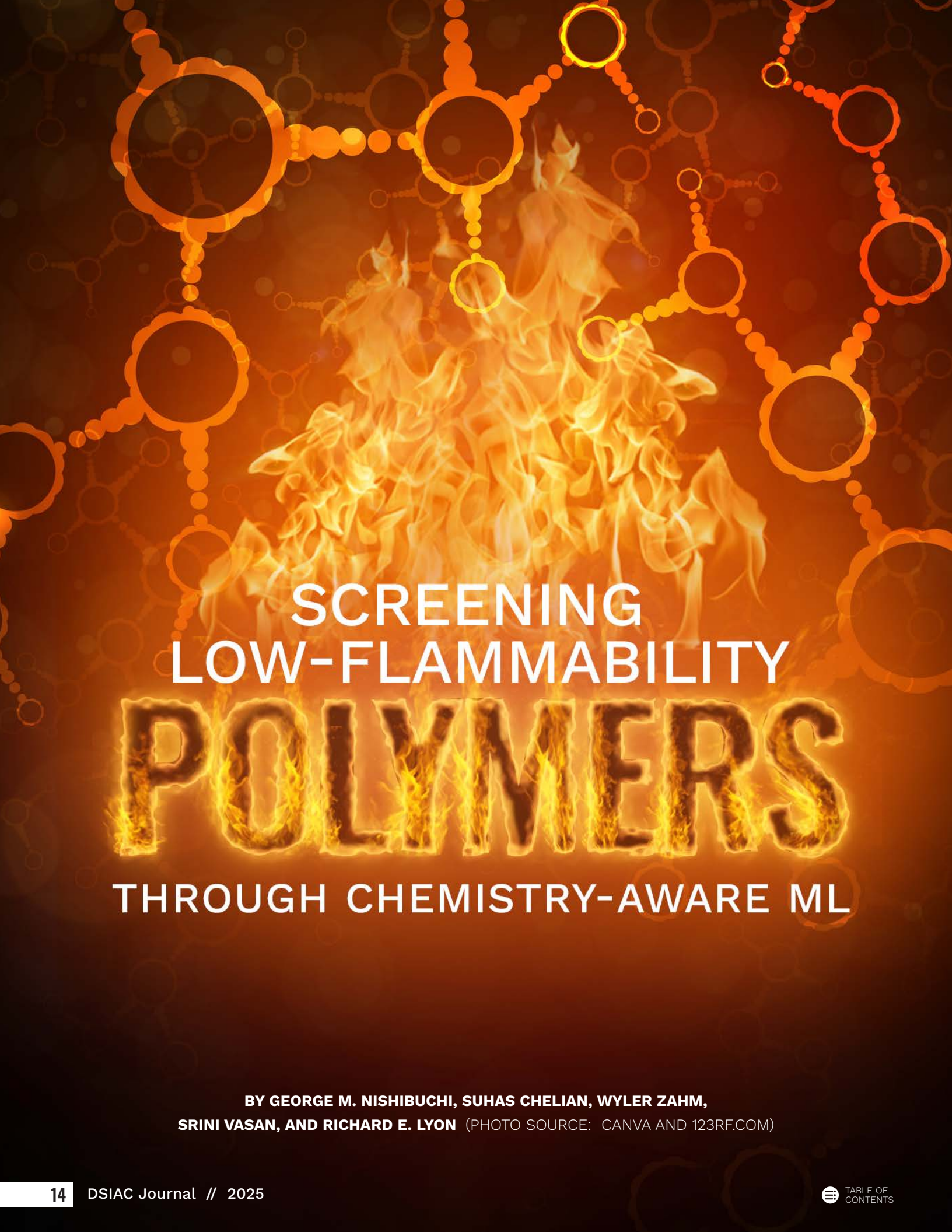
**JONATHAN HAASE** is the program manager of the U.S. Navy’s Expeditionary Missions program office (PMS 408), an acquisition and project management professional, an explosive ordnance disposal (EOD) officer, and a registered SCRUM trainer. His work at PMS 408 focuses on UUVs, remotely operated vehicles, and EOD equipment. He has also published on AI and decision making and metacognition and AI. Capt. Haase holds degrees from the U.S. Naval Academy, University of Maryland College Park, and professional certificates from Harvard.

# GIVE YOUR RESEARCH A BIGGER AUDIENCE



Photo Source: Getty Images  
Signature (Canva)

If you have research paid for by the DoD or U.S. government, contact us to get it published on DTIC's Research & Engineering Gateway. To learn more, visit <https://dsiac.dtic.mil/sti-collection>.



# SCREENING LOW-FLAMMABILITY POLYMERS THROUGH CHEMISTRY-AWARE ML

BY GEORGE M. NISHIBUCHI, SUHAS CHELIAN, WYLER ZAHM,  
SRINI VASAN, AND RICHARD E. LYON (PHOTO SOURCE: CANVA AND 123RF.COM)



# INTRODUCTION

**M**arine and aerospace vessels have stringent flammability requirements, defined through the American Society for Testing and Materials (ASTM) E1354 Cone Calorimetry standard [1], to protect those onboard Navy vessels in case of fire emergencies. The ASTM E1354 cone calorimeter experiment is a standardized method used to evaluate the fire behavior of materials by measuring parameters such as heat release rate, smoke production, and mass loss.

In this procedure, a small, square specimen is placed horizontally beneath a conical radiant heater, which emits a controlled heat flux, typically ranging from 10 to 100 kW/m<sup>2</sup>. An ignition source like a spark igniter is applied to initiate combustion once the specimen reaches its ignition temperature. Combustion gases are collected through an exhaust system, where oxygen consumption is measured. This allows calculating the heat release rate based on the principle that a known amount of energy is released per unit of oxygen consumed. Additional sensors monitor parameters like carbon monoxide and carbon dioxide production, smoke density, and mass loss rate to assess the material's flammability characteristics under controlled conditions.

The search for novel, low-flammability polymers has historically been an

experimentally intensive effort due to the large amounts of potential formulations, synthesis, and testing required for characterizing their flammability performance via the ASTM E1354 standard [1, 2]. Machine-learning (ML) and deep-learning (DL) methods have proven to be effective in screening molecules for properties unrelated to flammability but have not yet been used for predicting the ASTM E1354 standard properties based on the molecular structure of the polymers of interest [3, 4]. The search for novel polymer formulations with improved flammability performance begins with the molecular structure of the polymer; thus, there is a strong interest in being able to computationally predict their performance due to the extreme amount of potential polymer compositions.

“

***Machine-learning and deep-learning methods have proven to be effective in screening molecules for properties unrelated to flammability.***

This work combined atomistic density functional theory (DFT) simulations and cheminformatics to predict experimental cone calorimetry results through six different ML and DL models. By creating models based on molecular data, efficient screening of new potential polymer candidates by sidestepping the cost-intensive task

of experimental determination of polymer flammability properties for novel formulations with improved performance has been shown. The ML tasks involved predicting four variables defined in ASTM E1354—peak heat release rate (PHRR), average heat release rates (Avg. HRR) at 180 and 300 s, and time to ignition (TTI). An ensemble of six ML models was trained on a corpus of two experimental cone calorimetry databases using both features obtained from DFT simulations of the relevant polymers and those generated using cheminformatics methods.

Developing flame-resistant polymers is an ongoing necessity toward facilitating safe operating conditions for Navy and U.S. Department of Defense (DoD) personnel across all sectors.

## METHODS

ML/DL methods are as effective as the data that they train on. Thus, collecting and cleaning high-quality data are imperative to developing accurate models of flammability properties. In this section, the datasets used for training the ML/DL models in this work are described.

### Datasets

Two datasets were used for training ML models for predicting PHRR, TTI, and Avg. HRR at 180 and 300 s. They are summarized in the following subsections.

### **Federal Aviation Administration (FAA) Cone Calorimetry Database**

This experimental dataset collected by the FAA contained 211 full cone calorimetry experiments for various polymers—19 unique neat resins were observed in this dataset. Most of the polymers had multiple tests for each composition except for polyphthalamide, which only had one cone calorimeter experiment data available. The polymer having the most data available was the PT-30 phenolnovolac cyanate ester composition, with 28 experimental samples. With polymer compositions ranging from polyethylene (having the least desirable flammability properties) to those of the bisphenol C cyanate (the most desirable, having passed MIL-STD-2031 [5]), the dataset broadly covered the range of flammability properties seen in currently known polymer compositions.

### **Texas A&M University (TAMU) Dataset**

Prof. Wang of TAMU built a flame-retardancy database of more than 800 polymeric nanocomposites, including information from polymer flammability, thermal stability, and nanofiller properties [6]. This dataset was included to account for the varying types of fillers and additives that can affect flammability characteristics.

### **Synthetic Data Generation**

DL and ML models greatly benefit

from having a large amount of data to train on. This naturally conflicts with the high cost of performing a large amount of physical experiments to create data to train these models. Significant effort was put into hyperparameter tuning of the previously mentioned models to predict PHRR, TTI, and Avg. HRR at 180 and 300 s. In addition, studies on generating synthetic data were performed by using generative adversarial networks (GANs) specialized for generating data to mimic the distribution of the training data to smooth out the distribution of the dataset.

For each polymer composition, the monomer was converted to a simplified molecular-input line-entry system (SMILES) string representation. This was chosen, as most cheminformatics methods for feature generation were based on this representation. The Mordred library was chosen for feature generation, allowing the generation of ~2000 features for each molecule as a unique fingerprint to link to its flammability characteristics [6]. Such fingerprints are common in developing models that predict molecular properties and behaviors, such as biological activities and physical-chemical properties, which are fundamental in drug design and other chemical informatics applications. Mordred generates features either based on two- or three-dimensional representations of the molecule.

By providing these extensive and efficiently calculated descriptors, Mordred provided a wide range of features for this application, particularly since these features proved to be effective in quantitative structure-activity relationship and quantitative structure-property relationship modeling. Its ability to generate a comprehensive set of molecular descriptors, coupled with its open-source accessibility and ease of use, made it a good fit for this work.

### **Model Training**

Due to the limited amount of data available, a 95%–5% train-test split was chosen. Five-fold cross-validation was performed during training to mitigate overfitting. An ensemble of both classical and DL models was trained to compare their performance in predicting polymer flammability properties. These models were the Linear Regressor, Decision Tree Regressor, Deep Neural Network, Extreme Gradient Boosting (XGB) regressor (XGBoost), and Random Forest, as implemented in the Scikit Learn package. Specific implementation details for the algorithms are as follows:

- Two architectures for the deep neural network were explored—one with three hidden layers and one with six. For the network with three hidden layers, 256, 128, and 64 neurons were used for each layer. Rectified Linear Unit (ReLU) was



chosen as the activation function for the neurons. The Adam optimizer was used for training the model with a learning rate of .001, and training was performed for 500 epochs.

- The DL network with six layers was tested with 512, 256, 128, 64, 32, and 5 neurons per layer. ReLU was chosen as the activation function for the neurons. The Adam optimizer was used for training the model with a learning rate of .001 and was trained for 50 epochs.
- The XGBoost algorithm was trained with the Squared Error objective function. All other parameters were left as defaults.
- The random forest model had 100 splits, with all other parameters left as defaults.

A simplified overview of the models used in this project is provided in the Discussion and Results section.

## DFT Calculations

DFT calculations were performed using the CP2K simulation suite. Geometric optimization and electronic optimization calculations were performed for each of the unique polymer compositions in the FAA dataset. These simulations were performed using the Becke, three-parameter Lee-Yang-Parr (B3LYP) functional at the 6-31g\*\* (split-valence double-zeta basis set, including polarization functions) level of theory. Performance was tested

using central processing unit (CPU) and graphics processing unit (GPU) acceleration. Results showed that the boosts to performance on GPU were not relevant for the relatively small molecules studied. Therefore, these simulations were performed using an AMD EPYC 7413 CPU rather than an NVIDIA A100 GPU.

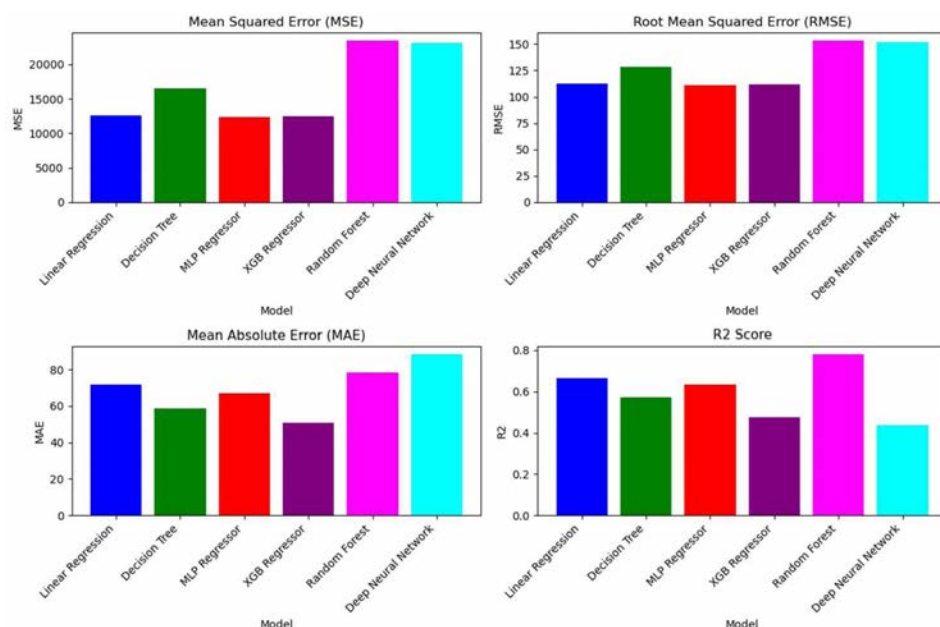
## DISCUSSION AND RESULTS

### Model Selection on the FAA Cone Calorimetry Dataset

After implementing these models, regression evaluation metrics were chosen to downselect across the models. The regression metrics included mean squared error (MSE), square root of MSE (RMSE), mean absolute error (MAE), and coefficient

of determination (R2). For MSE, RMSE, and MAE, lower values were desired. For an R2 score, higher was desirable. Metrics were averaged over several target variables, which were PHRR, average heat release rate (HRR) at 180 s, average HRR at 300 s, time to sustained ignition, and average specific extinction area.

Broadly, the limitations in the amount of data available significantly hindered the accuracy of the model across all metrics, despite the use of molecular fingerprints. Using MAE as a target metric, XGB and Random Forest performed best. Using R2 as a target metric, Random Forest and Linear Regression performed best. These results are shown in Figure 1, where comparing ML models with different metrics averaged over several target variables on the FAA data. R2 was 0.7 when averaging over several factors



**Figure 1.** Predictions on the FAA Dataset (Source: G. M. Nishibuchi).

like polymer type; external heat flux and metrics such as peak heat release, time to ignition and char formation; and up to R2 of 0.93 for XGB on an average heat release at 300 s.

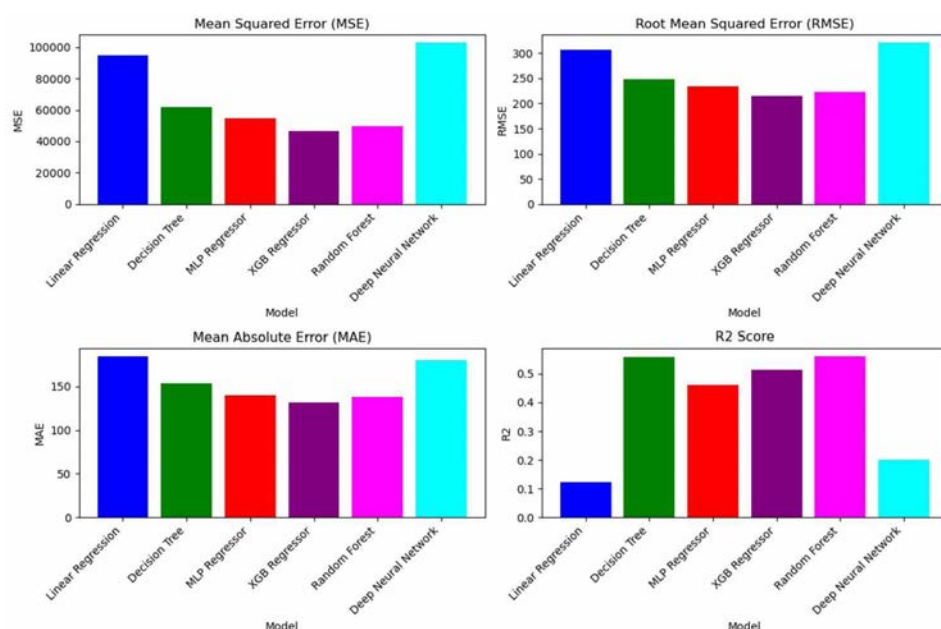
Decision trees, Random Forest, and XGB are similar algorithms, and XGB is often used in literature. For this reason, experiments with XGB continued in this analysis.

## Feature Selection: Training With and Without DFT Features

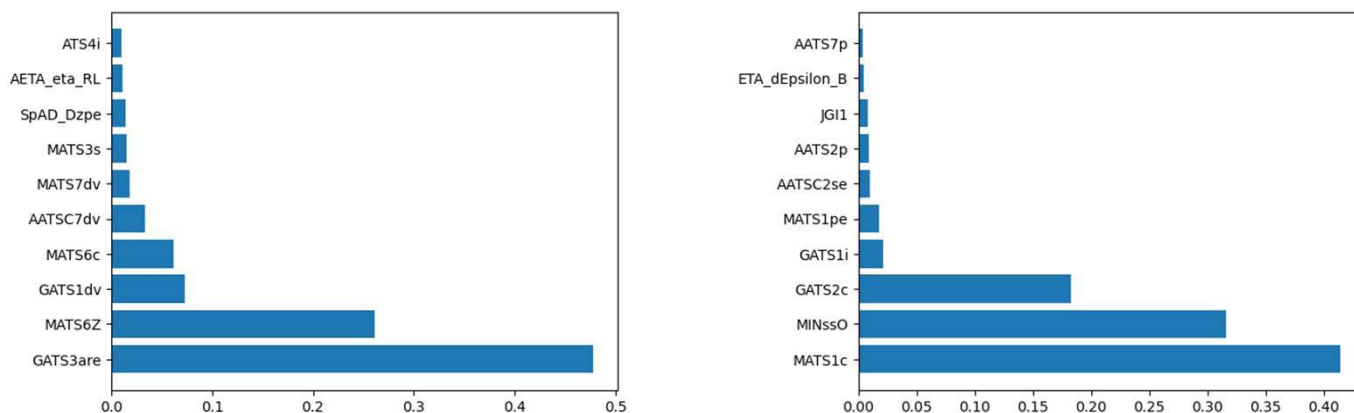
Limitations were necessary on the amount of data that could be obtained from the DFT calculations regarding whether the highest energy occupied molecular orbital (HOMO)-lowest energy unoccupied molecular orbital (LUMO) gap and free energies from the DFT calculations made a meaningful impact on model performance. Thus, model training was performed with and without DFT features. With the DFT features included, MSE, RMSE, and MAE all rose and R2 scores fell (Figure 2).

Tree-based models like XGBoost can generate feature importance values, which track how many times a feature is used to split a node in the decision tree. Each split reduces the Gini impurity, thus reducing the likelihood of the model selecting a random point in the dataset. The features with the highest amount of attributed splits can be considered the most important toward making a correct prediction.

Figure 3 shows the 10 highest importance Mordred features for PHRR (left) and TTI (right). Mordred features can be difficult to describe succinctly (e.g., GATS3are is the Geary coefficient of lag 3 weighted by Allred-Rochow electronegativity) but are understood by computational chemists and repeatable.



**Figure 2.** Comparison of Models Averaged Over Several Target Variables on the FAA Data (Source: G. M. Nishibuchi).



**Figure 3.** Feature Importance Across a Subset of Mordred Features for PHRR (Left) and TTI (Right) (Source: G. M. Nishibuchi).

## Synthetic Data Generation Effects on Prediction Metrics

GANs are ML algorithms that generate synthetic data mimicking real data. A GAN consists of two neural networks—the generator and the discriminator. The generator creates fake data from random noise, aiming to produce data indistinguishable from real data. The discriminator evaluates both real and fake data, attempting to distinguish between them. The generator and discriminator are trained simultaneously in an adversarial process. The generator improves by creating more realistic data to fool the discriminator, while the discriminator enhances its ability to detect fake data. GANs have been applied successfully in various fields, including image and video generation, text creation, and data augmentation, due to their ability to generate high-quality synthetic data. While stable diffusion has generally taken over the image generation space (e.g., the DALL-E series of models), GANs have proven effective in the domain of generating tabular data.

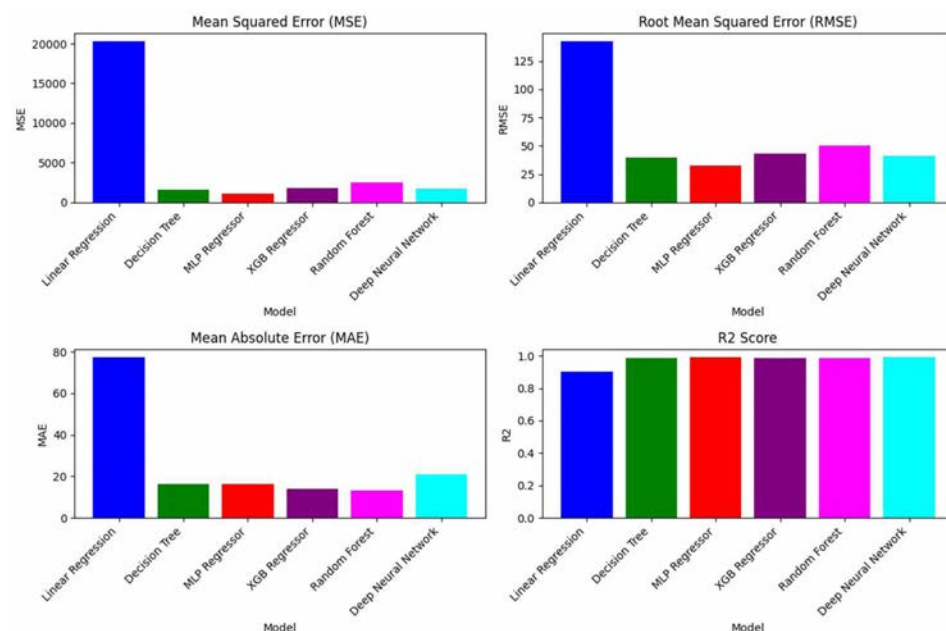
A variety of synthetic tabular data generation models was tested, including variational autoencoders, diffusion models, CTGANs, triplet-based VAE, bootstrapping, and a Gaussian copula synthesizer. A dataset containing ~36,000 samples was created based off the data contained in the FAA database and significantly improved the performance across

multiple ML and DL models, with metrics shown in Figure 4.

### Composite Prediction

Composites of multiple polymers are a particularly difficult problem for predicting polymer flammability due to the limited amount of data existing for the neat resins alone. However, composites often provide favorable material properties compared to the properties of the individual neat resins; thus, it is necessary to predict the flammability characteristics of polymer composite systems. A methodology for predicting the properties of polymers was formulated as a weighted sum of the properties of the components of a composite or a polymer with additives. The calculation is formalized as follows:

$$P_{\text{composite}} = \sum_i P_{\text{resin}_i} * W_i + \sum_j P_{\text{FR}_j} * W_j, \quad (1)$$



**Figure 4.** Comparison of Model Metrics for Predicting PHRR With (Top) and Without (Bottom) Synthetic Data (Source: G. M. Nishibuchi).

“

*Composites of multiple polymers are a particularly difficult problem for predicting polymer flammability due to the limited amount of data existing for the neat resins alone.*

where  $P_{\text{composite}}$  is the composite property of interest (PHRR, TTI, etc.),  $P_{\text{resin}_i}$  is the predicted property of the individual polymer component,  $W_i$  is the mass percentage of the polymer component,  $P_{\text{FR}_j}$  is the predicted property for a flame retardant/additive, and  $W_j$  is the mass percentage of the flame retardant/additive. This allows



the user to get a sense of the range of potential predictions and decide which values are best to use/not use (e.g., in the case of an obvious outlier). This also helps in identifying which models struggle with predicting certain parameters so the user can choose to retrain the models on different/ altered data to improve performance or choose a different model for predictions.

## Study on Polymers With Flame Retardants (FRs) and Fillers

TAMU's Prof. Wang developed a dataset of polymer flammability before and after the addition of FRs. The TAMU dataset was used to predict the flame retardancy index (FRI), TTI, total heat release (THR), and PHRR given several input features. Raw features and custom features were combined as input variables to improve model performance. The features in the dataset are explained as follows:

- **Flammability:** Baseline flammability of the pure polymer.
- **TGAP:** Thermal stability of the pure polymer.
- **Nanofiller loading (wt):** Amount of nanofiller added.
- **dTGA:** Change in thermal stability due to nanofiller.
- **Dimension:** Shape and form factor of the nanofiller.
- **Type:** Material composition of the nanofiller.

- **IFR:** Presence of intumescent flame retardant.

The custom features are given as follows:

- **Average TGA:** Average of the TGA values before and after flame-retardant treatment.
- **Polymer and Type:** Combines the polymer type and another categorical feature and type into a single feature, creating a combined categorical feature that uniquely identifies the combination of polymer type and another characteristic.
- **IFR Flammability Ratio:** Ratio of flammability to the presence of an intumescent flame retardant. The ratio aims to quantify the flammability relative to the presence of an intumescent flame retardant. Adding 1 to the IFR value ensures the denominator is never zero, preventing division errors. This feature can highlight how flammability changes with and without the flame retardant.
- **Flammability:** Represents the flammability measurement of the sample.
- **IFR:** A binary feature indicating the presence (1) or absence (0) of an intumescent flame retardant.
- **Polymer and Incorporated Nanoparticles:** Combines the polymer type and the type of incorporated nanoparticles into a single feature, creating a combined categorical feature that uniquely

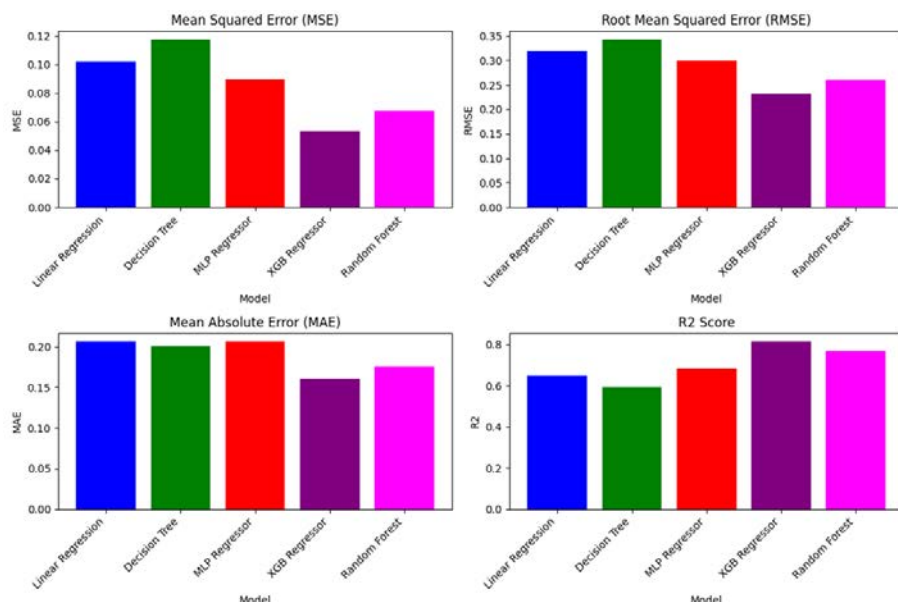
identifies the combination of polymer type and incorporated nanoparticles. This can capture the specific interactions between the polymer and the type of nanoparticles used, which could be relevant for the model.

Using these input variables, models were trained to predict the following target variables: After FR TTI, PHRR, THR, and FRI. Regression metrics across the trained models and averaged over several target variables on the TAMU dataset are shown in Figure 5. Performance was acceptable, with up to a 0.8 R<sup>2</sup>.

XGB outperformed all other models in terms of MSE, RMSE, MAE, and R<sup>2</sup>. As with the FAA dataset, feature importance was also examined. Using a Random Forest classifier, which is similar to XGB, the feature importance for each target variable was shown. Before FR TTI, PHRR and THR were important. In incorporated nanoparticles, percent weight was also important. Feature importance values from the XGBoost model are shown in Figure 6 for the four target variables.

“

*XGB outperformed all other models in terms of MSE, RMSE, MAE, and R<sup>2</sup>.*



**Figure 5.** Comparison of ML Models With Different Metrics (Source: G. M. Nishibuchi).

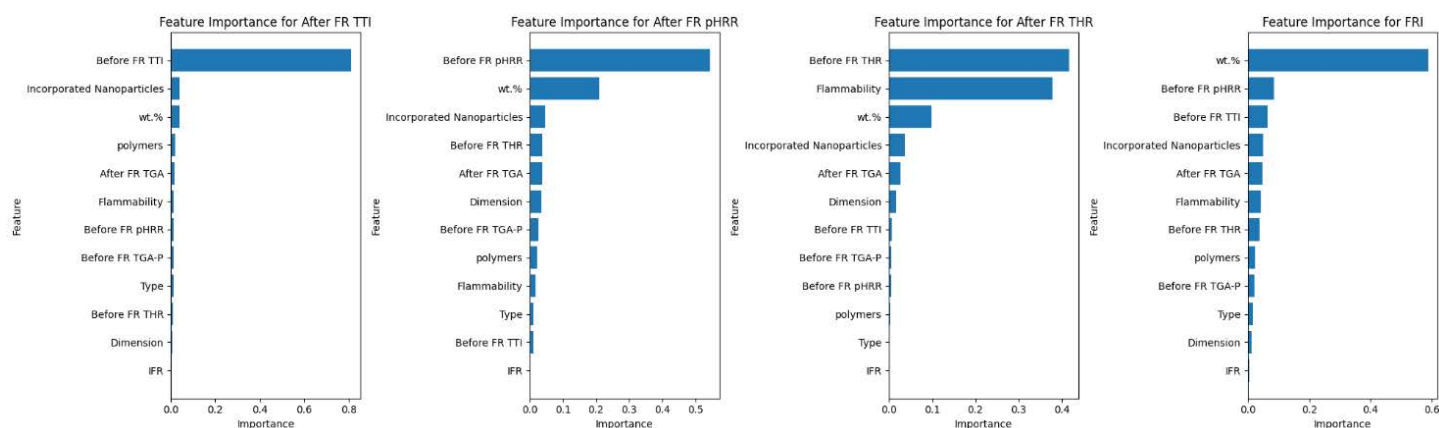
## DFT Simulations

It was found that the CP2K software suite was better optimized for the CPU than the GPU, with optimization iteration times being an order of magnitude lower than on the GPU. DFT software applications are generally written using Fortran due to its high efficiency and extremely well-established set of linear optimization and matrix multiplication libraries that

are core parts of any DFT software. Because of the high complexity of these software applications, they have also been slow to uptake modern GPUs designed for the matrix operations used in linear optimization and matrix multiplication. The poor GPU performance relative to CPU observed in CP2K is a possible result of years of CPU optimization being compared to a relatively new and

unoptimized GPU implementation. Another key contributor to this issue is the fact that the monomer system sizes are likely not large enough to benefit from the use of massively parallelized GPUs.

A study from Los Alamos National Laboratory that involved performance benchmarking comparisons between CPU- and GPU-based CP2K implementations showed up to a 3.7× boost in performance on a GPU system compared to an identical simulation on a CPU [7]. Their system size was on the order of 900 atoms, and they did observe a significant boost in performance at the much larger system size. It is highly likely that small system sizes (<50 atoms) under study did not necessitate the high-throughput advantages of the GPU, and the increased amount of overhead from GPU-CPU memory transfers led to longer runtimes compared to the CPU-only implementation.

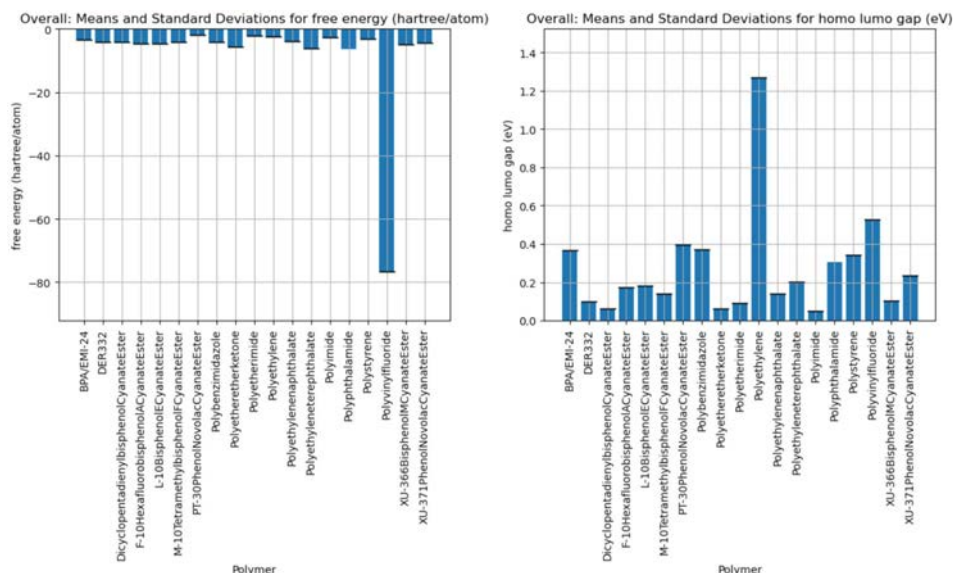


**Figure 6.** Top Feature Importance Values for Different Target Variables From the XGBoost Model (Source: G. M. Nishibuchi).

In Figure 7, one incorrect calculation was observed for the free energy of polyvinyl fluoride due to an issue with the geometric optimization's convergence. While polyethylene's HOMO-LUMO gap appears to be an outlier, it was evaluated against literature values and found to be consistent. In a larger system, this gap would decrease because of conjugated bonds in the larger polymer chain.

## CONCLUSIONS

Predicting macroscale cone calorimeter measurements from atomistic features comes with a wide variety of challenges, both theoretical and practical. This work demonstrated the development of molecular fingerprints for predicting polymer flammability, first principles simulations of neat resins, and the development of ML and DL molecules for predicting ASTM E1354 experimental measurements from DFT, experimental, and molecular properties. The use of ensemble prediction was found to be effective in the low data domain of ASTM E1354 experimentation. Continued work in developing standardized databases for storing cone calorimetry data is imperative for further development of ML methods to predict polymer flammability. ■



**Figure 7.** Free Energy (Left) and HOMO-LUMO Gap (Right) for the Unique Polymer Compositions in the FAA Database, Obtained via DFT Simulations (Source: G. M. Nishibuchi).

## ACKNOWLEDGMENTS

This material was based upon work supported by the U.S. Office of Naval Research under award number N6833524C0125.

## REFERENCES

- [1] Lyon, R. E., et al. "A Statistical Model for the Results of Flammability Tests." Conference Proceedings – Fire and Materials, The 11th International Conference and Exhibition, 2009.
- [2] Hergenrother, P., et al. "Flame Retardant Aircraft Epoxy Resins Containing Phosphorus." *Polymer*, vol. 46, no. 14, pp. 5012–5024, <https://doi.org/10.1016/j.polymer.2005.04.025>, 2005.
- [3] Ma, T., et al. "Thermal Degradation and Flame Retardancy Prediction of Fe, Al, and Cu-Based Metal-Organic Framework and Polyethylene Terephthalate Nanocomposites Using DFT Calculation." *Polymer*, vol. 263, p. 125496, <https://doi.org/10.1016/j.polymer.2022.125496>, 2022.
- [4] Nguyen, H. T., et al. "Predicting Heat Release Properties of Flammable Fiber-Polymer Laminates Using Artificial Neural Networks." *Composites Science and Technology*, vol. 215, p. 109007, <https://doi.org/10.1016/j.compscitech.2021.109007>, 2021.
- [5] U.S. DoD. *Fire and Toxicity Test Methods and Qualification Procedure for Composite Material Systems Used in Hull, Machinery, and Structural Applications*

*Inside Naval Submarines.* MIL-STD-2031, 26 February 1991.

- [6] Moriwaki, H., Y.-S. Tian, N. Kawashita, and T. Takagi. "Mordred: A Molecular Descriptor Calculator." *Journal of Cheminformatics*, vol. 10, no. 4, doi: 10.1186/s13321-018-0258-y, 2018.
- [7] Yokelson, D., N. V. Tkachenko, R. Robey, Y. W. Li, and P. A. Dub. "Performance Analysis of CP2K Code for Ab Initio Molecular Dynamics on CPUs and GPUs." *Journal of Chemical Information and Modeling*, vol. 62, no. 10, pp. 2378–2386, <https://doi.org/10.1021/acs.jcim.1c01538>, 2022.

## BIBLIOGRAPHY

- Glassman, I., and R. A. Yetter. *Combustion*. Academic Press, 2008.
- Goodstein, D. L. *States of Matter*. Dover Publications, Inc., 2017.
- Lyon, R. E., et al. "A Molecular Basis for Polymer Flammability." *Polymer*, vol. 50, no. 12, pp. 2608–2617, <https://doi.org/10.1016/j.polymer.2009.03.047>, 2009.
- Lyon, R. E., et al. "Fire-Resistant Aluminosilicate Composites." *Fire Mater.*, vol. 21, pp. 67–73, [https://doi.org/10.1002/\(SICI\)1099-1018\(199703\)21:2<67::AID-FAM596>3.0.CO;2-N](https://doi.org/10.1002/(SICI)1099-1018(199703)21:2<67::AID-FAM596>3.0.CO;2-N), 1997.
- Pomázi, A., et al. "Predicting the Flammability of Epoxy Resins From Their Structure and Small-Scale Test Results Using an Artificial Neural Network Model." *Journal of Thermal Analysis and*



*Calorimetry*, vol. 148, no. 2, pp. 243–256, <https://doi.org/10.1007/s10973-022-11638-4>, 2022.

Quan, Y., Z. Zhang, R. N. Tanchak, et al. “A Review on Cone Calorimeter for Assessment of Flame-Retarded Polymer Composites.” *Journal of Thermal Analysis and Calorimetry*, vol. 147, pp. 10209–10234, <https://doi.org/10.1007/s10973-022-11279-7>, 2022.

Stoliarov, S. I., et al. “Prediction of the Burning Rates of Non-Charring Polymers.” *Combustion and Flame*, vol. 156, no. 5, pp. 1068–1083, <https://doi.org/10.1016/j.combustflame.2008.11.010>, 2009.

Tao, Q., et al. “Machine Learning for Perovskite Materials Design and Discovery.” *Computational Materials*, vol. 7, no. 1, npj, <https://doi.org/10.1038/s41524-021-00495-8>, 2021.

## BIOGRAPHIES

**GEORGE M. NISHIBUCHI** is a senior ML engineer and computational materials scientist at Quantum Ventura Inc. He has run over 50,000 DFT simulations as a researcher at Purdue University’s

Network for Computational Nanotechnology, from high-throughput studies of semiconductors to mechanistic studies in solid-state electrolytes. Mr. Nishibuchi holds a B.S. and M.S. in materials engineering from Purdue University.

**SUHAS CHELIAN** is a researcher and ML engineer at Quantum Ventura Inc. He has captured and executed more than \$12 million worth of projects with several organizations like Fujitsu Labs of America, Toyota (Partner Robotics Group), Hughes Research Lab, the Defense Advanced Research Projects Agency, the Intelligence Advanced Research Projects Agency, and the National Aeronautics and Space Administration. He has 31 publications and 32 patents demonstrating his expertise in ML, computer vision, and neuroscience. Dr. Chelian holds dual bachelor’s degrees in computer science and cognitive science from the University of California San Diego and a Ph.D. in computational neuroscience from Boston University.

**WYLER ZAHM** is a senior ML engineer at Quantum Ventura Inc. He has worked with advanced algorithms, front- and back-end development, a variety of artificial intelligence (AI)/ML architectures and frameworks like full precision/GPU and reduced precision/neuromorphic technologies, and applications like automated vulnerability

detection and repair for computer source code and cybersecurity. Mr. Zahm has dual bachelor’s degrees in computer engineering and data science from the University of Michigan.

**SRINI VASAN** is the president and chief executive officer at Quantum Ventura Inc. and chief technology officer at QuantumX, the research and development arm of Quantum Ventura Inc. He specializes in AI/ML, AI verification and validation, ML quality assurance and rigorous testing, ML performance measurement, and system software engineering and system internals. Mr. Vasan studied management at the MIT Sloan School of Management.

**RICHARD E. LYON** is program director at the FAA. He has invented many of the techniques used for polymer and composite materials analysis regarding flammability and polymer composites by hand for the Navy. He has written landmark papers on the study of polymers, composites, and their properties, including relevant experiments at the Lawrence Livermore National Laboratory and the FAA. Dr. Lyon holds a Ph.D. in polymer science and engineering from the University of Massachusetts Amherst.

# WANT TO READ MORE?

If you found this publication insightful and engaging, please check out our back issues on <https://dsiac.dtic.mil>. We also offer similar journals covering the cybersecurity and homeland defense and security spheres, which you can find at <https://csiac.dtic.mil> and <https://hdiaac.dtic.mil>.









# DATA-OPTIMIZED HUMAN- MACHINE TEAMING WITH **ROBOTIC WINGMEN**

BY CHRISTINA HAYHURST, CHRISTINE COVAS-  
SMITH, AND PATRICIA HARRIS (SOURCE: CANVA AND  
ADOBE STOCK)

## INTRODUCTION

**T**he 2022 National Defense Strategy directs the U.S. Department of Defense (DoD) to urgently act to strengthen the U.S. military against its pacing challenge—the People’s Republic of China (PRC) [1]. The PRC challenges the U.S. military’s information advantage in the operational environment, undermining kinetic maneuver across all domains of the Joint Force [2]. Within the air domain, the PRC’s ability to disrupt, deny, and degrade the data informing air operations will threaten the U.S. Air Force’s Operational Imperative of tactical air dominance and the Air Force Future Operating Concept of the



successful fight for air superiority at the forward edge of the battlespace. Human-machine teaming (HMT) with artificial intelligence (AI)-driven air power platforms, such as the uncrewed, autonomous, “robotic wingmen” aircraft known as collaborative combat aircraft (CCA), will thwart the PRC’s pursuit of information advantage. CCA employment will shrink the kill chain and expand the air domain’s lethality by providing qualitative and quantitative optimization of Air Force operators’ observe, orient, decide, and act (OODA) loop and targeting cycle.

However, HMT optimization depends upon human-data accountability, which is the human ownership, understanding, and implementation of the data that the CCAs use to execute tactical decisions autonomously. In the future operating environment, data will be the critical pivot between human command of CCAs and CCA algorithmic effects on the battlespace. To best equip the Joint Force for future CCA employment against the PRC threat, the DoD must consider new criteria to determine the next generation of its readiness reporting for robotic wingmen—the combat readiness of CCA data.

This article will recommend a CCA data readiness tool that quickly informs commanders of their level of risk assumption based on data criteria rankings and using the tool’s application to cognitive electronic warfare (CEW) CCAs as a vignette.

“

***In the future operating environment, data will be the critical pivot between human command of CCAs and CCA algorithmic effects on the battlespace.***

It will also suggest the personnel, policy, and technology investments to optimize each data readiness criterion.

## **AUTONOMOUS WEAPONS SYSTEM’S (AWS’s) “BLACK BOX” CHALLENGE**

AWSs are transformational technologies for future warfare. The incremental development of AWSs that “once activated, can select and engage targets without further human intervention” has led to human operators moving further from the immediate decision-making on the use of force [3]. This human-machine interaction weaponizes AI by distributing an agency to an AWS that is inaccessible to human reasoning. While many computational techniques are summarized under the “AI” term, autonomous systems that the military are currently developing fall under the “narrow AI” category.

At the heart of narrow AI applications are machine-learning (ML) algorithms

that are data-hungry and data-dependent [3]. AWS’s reliance on data processed by ML algorithms presents a fundamental problem for commanders using it on the battlefield. Today’s ML algorithms are in a black box that cannot explain or guarantee certain behaviors. This problem raises an important question—how can commanders be reasonably responsible for using an AWS if they do not know the system’s decision-making process? Fortunately, AWS’s current black box behavior is not a foregone conclusion that commanders must reluctantly accept as necessary for future Warfighting and information dominance.

The question of trust in machines for risk calibration and the extent of meaningful human control (MHC) for the ethical employment of autonomous weapons systems are not novel. As weapons systems have become technologically sophisticated, research regarding machine trust and MHC has grown in practical application. The lessons gained from past automatic weapon employment and MHC research form a valuable foundation to consider future commander trust and risk assessment of autonomous systems on the battlefield.

## **MHC APPLIED TO AWSs**

Autonomy and automation have “long been integrated into the critical functions of air defense systems to detect, track, prioritize,

select, and potentially engage incoming air threats” [3]. In their “automatic” mode, air defense systems autonomously deploy countermeasures if they detect a threat; however, human operators are “on-the-loop,” allowing them to supervise the system’s actions and abort the attack. In this case, human operators retain situational awareness and have sufficient insights into the parameters under which the command module selects and prioritizes targets. However, to break down the noteworthy problems with autonomous air defense operations, it is helpful to understand three dimensions of MHC discovered through human factors research—a technological dimension through weapon design, a conditional dimension that limits weapons use, and a decision-making dimension that defines acceptable human-machine interaction. All three dimensions must be considered when employing autonomous systems to reach an ethically responsible level of MHC.

In the case of air defense systems, the compromise of MHC has led to many severe incidents of friendly fire, specifically in the human-machine interaction dimension. For example, a series of fratricides involving the Patriot system, a human-in-the-loop air defense system, attributed to excess trust that made the system a de facto fully autonomous weapon [4]. A thorough analysis of these friendly fire incidents by autonomous air defense systems

identified the following challenges: automation bias or overtrust, lack of system understanding, lack of situational awareness, lack of time for deliberation, lack of human expertise, inadequate training, and operating under high-pressure combat situations [3].

As autonomy increases, the loss of user alertness is proportional to the system’s enhanced automation and perceived reliability, leading to the “automation conundrum” [4]. Despite this identified issue of MHC over Patriot equipment in its “automatic” mode during the fratricide incidents, the Army’s readiness assessment of Patriot units continues to be exclusively tied to its maintenance requirements and equipment replacement rates as part of its Patriot recapitalization program [5]. Autonomous operations have unfortunately increased without proper risk considerations and readiness evaluations of its data-driven autonomy algorithms, leading to operations with meaningless human control as an unfortunate yet still appropriate use of force.

## PAIRING HMT WITH MHC

This state of reluctant risk acceptance is not only incompatible with the future ethical employment of AWS but challenges optimal HMT during operational employment. The rise in automation necessitates a

reconceptualization of trust between humans and automated systems. Undertrust in a system can lead to its lack of use, and overtrust can lead to complacency and poor monitoring [6]. Since under- and overtrust are problematic, appropriate trust calibration is critical to effective HMT and risk assessment. Increasing automation has led to the advent of a new HMT paradigm [7]. Within this paradigm, the machine is a teammate of the human, who has innovative abilities to be exploited rather than liabilities for which to be compensated. To determine how humans can appropriately calibrate trust in future AWSs to best facilitate HMT, it is helpful to examine how to apply MHC models to commander trust calibration in operational weapons systems with autonomous features [3].

Human factors research has shown three key variables influencing HMT with automated systems: (1) the human trustor, (2) the automated machine trustee, and (3) the context in which the interaction occurs [6]. These variables nest within MHC’s three technological, conditional,

“

*The rise in automation necessitates a reconceptualization of trust between humans and automated systems.*

and human-machine interaction dimensions in autonomous weapons systems. The technological dimension of MHC represents the machine's system factors, which include physical system attributes and performance factors. The conditional dimension of MHC includes environment and context-related factors, which involve team collaboration and task-based factors like type and complexity. MHC's final human-machine interaction dimension represents the human trustor, which clarifies the human's understanding of his or her role in the shared work with an automated system.

For human operators to regain meaningful control of autonomous systems, which enables the appropriate calibration of the trust in employing combat-ready robotic wingmen, the following three prerequisite conditions must be met that apply human factors variables and MHC dimensions to CCAs [3]:

1. A functional understanding of how the targeting system operates (automated machine trustee variable with the technological dimension).
2. Sufficient situational understanding (context variable with the conditional dimension).
3. The capacity to scrutinize machine targeting decision-making (human trustor variable with the human-machine interaction dimension).

Optimized teaming between Warfighters and AWSs begins with optimized human-data accountability that hinges on these three HMT human factors variables combined with the three dimensions of MHC—technological, conditional, and human-machine interaction.

## APPLYING CCA WITH THE CEW MISSION

Within the Air Force, CCAs will be the future air domain AWS that flies autonomously alongside crewed platforms, testing HMT concepts against the PRC threat [8]. CCAs will harness autonomy, AI, and ML to present formidable airpower capacity against hostile air threats in highly contested environments. These loyal, robotic wingmen can team with human operators by offloading data analysis tasks such as suggesting flight corridors, mapping targets, and appropriate courses of action [4]. Additionally, to increase the survivability of crewed platforms within the PRC's highly contested and lethal environment, dense with anti-access/area denial (A2/AD) capabilities, CCAs can also saturate China's People's Liberation Army (PLA) defenses or autonomously deliver kinetic effects.

Although CCAs will be designed to perform various mission sets, this article will recommend CCA data readiness criteria for evaluating



***CCAs will harness autonomy, AI, and ML to present formidable airpower capacity against hostile air threats in highly contested environments.***

cognitive electronic warfare (CEW)-designated CCAs. It will use tactical, forward-edge CEW CCAs as a case study to address the most fundamental threat CCAs could face to challenge optimized HMT—effects against its data. Electronic warfare (EW) uses the electromagnetic spectrum (EMS) to deliver effects against the enemy's use of the EMS. The next generation of EW is CEW weapons systems that use AI and ML to automate EW decisions involved in the detection, signal classification, prediction of enemy EW tactics, and countermeasure execution. Since freedom of EMS maneuver provides kinetic maneuver, future wars will be won or lost in the EMS, making CEW CCAs critical in potential conflict against the PRC.

Until recently, EMS threats did not change quickly, so the EW integrated reprogramming (EWIR) process could take months to reconfigure operational flight programs [9]. However, PRC EW assets have rapidly advanced, and responding to these assets requires faster updates than the EWIR



enterprise can accomplish. The PLA’s strategists insist on establishing EMS dominance through EW against U.S. assets through the deception strategy of “hide the real and inject the false,” affecting data to mislead U.S. operators [10].

Consequently, the new CEW CCA capability that the United States is developing must have data that is accessible, secure, and appropriately configured to deliver intended battlefield effects. Despite this need, commanders lack a readiness reporting tool to assess a CEW CCA’s readiness to fulfill its intended capability based on the combat readiness of the data driving its behavior. Furthermore, commanders have neither a means of qualifying how much risk they assume by employing robotic wingmen nor a means of ethical accountability for their CCA employment decision.

The operational concept of combat-ready data must be crafted to optimize HMT with future CEW CCAs to rapidly generate EW effects at the forward edge of the battlespace. Since Joint Force commanders currently lack the ability to appropriately determine if data that will drive CEW CCAs can achieve their intended effects of denying adversary objectives without being compromised or introducing ambiguity to friendly forces, the following seven criteria should be used to evaluate the combat readiness of data using both MHC dimensions and HMT factors:

1. Data security within the technological dimension.
2. Data trust within the technological dimension.
3. Data architecture within the technological dimension.
4. Data understanding within the conditional dimension.
5. Data accessibility within the human-machine interaction dimension.
6. Data visibility within the human-machine interaction dimension.
7. Data interoperability within the human-machine interaction dimension.

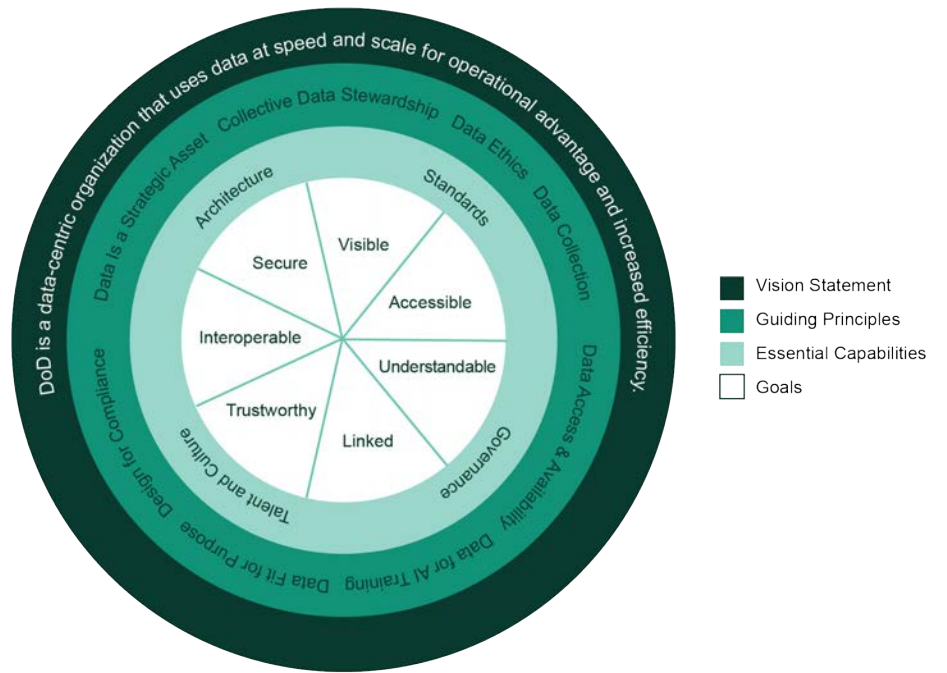
These seven criteria are not only the enabling objectives of the 2022 DoD Data Strategy, as illustrated in Figure 1, but they can also be uniquely

understood in their application to data-driven CEW CCAs [11].

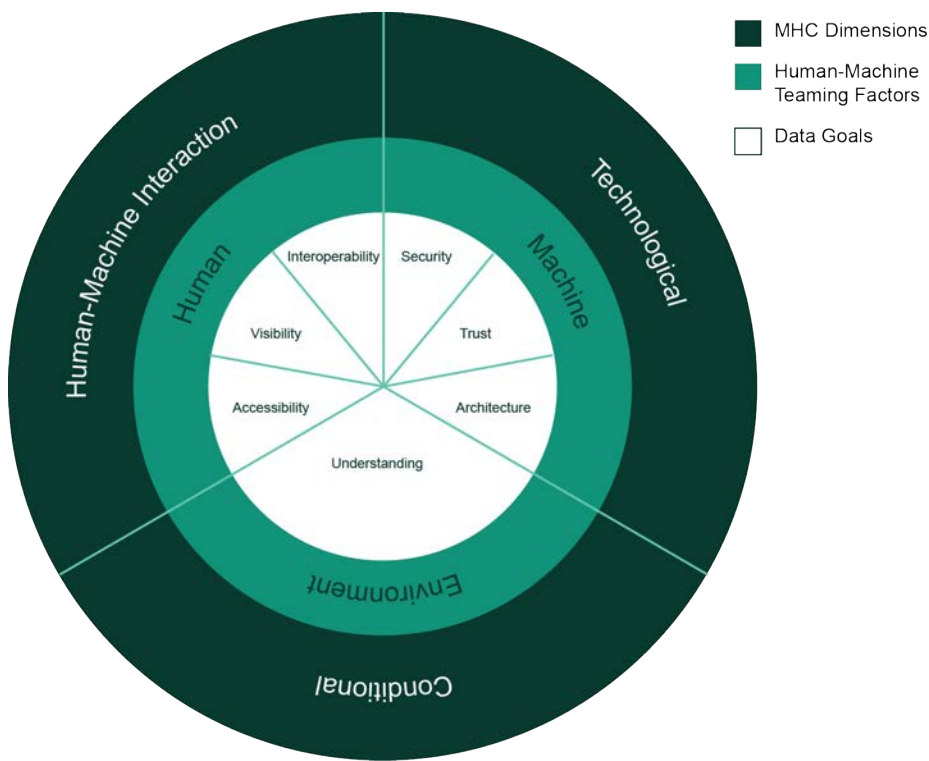
Figure 2 shows the proposed interrelationship between the MHC dimensions for ethical AWS employment, HMT factors for optimal AWS operations, and the 2022 DoD Data Strategy’s data goals for CCA data readiness. These concepts must be applied to future CCA use to ensure commanders are ethically accountable and risk-informed regarding CCA employment’s cost to achieving military objectives [12].

## DIMENSIONS OF MHC

Three dimensions characterize MHC—technological, conditional, and human-machine interaction. Figure 2 highlights how specific HMT factors



**Figure 1.** DoD Data Strategy of Vision, Principles, Capabilities, and Goals (Source: U.S. DoD [11]).



**Figure 2.** Interrelationship Diagram of MHC, HMT, and Data Goals (Source: C. Hayhurst, C. Covas-Smith, and P. Harris).

and CCA data readiness goals from the 2022 DoD Data Strategy nest within these three dimensions.

## Technological

The first dimension of MHC is the technological dimension, defined by data security, trust, and architecture. These factors are critical to characterizing the machine variable influencing human trust in automated systems.

### Data Security

Within the 2020 DoD Data Strategy, the DoD's Chief Data Officer (CDO) governs the DoD's data management efforts to ensure data security standards are met across the entire

department [11]. The data security approach recommended in the 2020 DoD Data Strategy is granular privilege management, which uses identity, attributes, and permissions to govern access to data through public key cryptography. However, once cryptanalytically relevant quantum computer capabilities are available, public key algorithms will be vulnerable to adversary attacks [13].

Quantum computing systems threaten current encryption mechanisms that provide the basis of internet commerce and communication. The PRC has surged in its quantum technology research and investments. Chinese companies dominate in quantum cryptography patents, and China

has taken the lead in the largest demonstrated network with quantum key distribution [14]. If nothing is done now to protect data streams, any encrypted data the PLA intercepts will be vulnerable to decryption in the future. CCAs with data protected by post-quantum cryptography (PQC) provide the best data security assurance [15]. Consequently, CCA data characterized by PQC security standards gives a commander the lowest risk to the mission within the data security criterion.

### Data Trust

Data trust is the next factor that informs the technological dimension of MHC of CCAs. The 2020 DoD Data Strategy describes trustworthy data as having proper tags and pedigree metadata throughout its life cycle [11]. These actions build user confidence in the data due to enhanced data quality, which is critical to support operational decision-making. The Chief Digital and Artificial Intelligence Officer (CDAO) further clarifies metadata governance in the 2023 DoD Metadata Guidance [16]. The guidance assumes that DoD organizations "will apply metadata at the most appropriate time between creation and storage and maintain the tagging through the data assets' life cycles" [16]. Trustworthy data is tagged data that supports the metadata functions of search and discovery, access control, correlation, audit, records management, and protection. CCAs that use data in

compliance with the DoD Metadata Guidance give commanders the lowest risk in data trust.

### **Data Architecture**

The third element of the technological dimension is data architecture. For CEW CCAs, a robust data architecture would enable the rapid assessment of CCA employment to determine the effectiveness of its adversary signal translations and countermeasure decisions. The data architecture that the Air Force will field that satisfies this element is the Advanced Battle Management System (ABMS), which allows data to be shared across multiple platforms as part of the DoD's Joint All Domain Command and Control effort [17].

The success of the ABMS data architecture depends on rigorous adherence to data standards that provide a common application environment and a set of flexible protocols [18]. This architecture would provide CCAs and their human operators with a totality of data to be used at the strategic level and tactical edge so the commander's intent can be met with relevant data at the location most applicable to CCA operation. Furthermore, a robust ABMS architecture would sense and synthesize data through AI/ML-based analytics at expected intervals, provisioning CCAs with as relevant and accurate a data threat picture as possible before reaching degraded or denied A2/AD environments.

To improve the speed and quality of its own information processing, the PLA is also pursuing a "system of systems" network under its informatized and intelligentized warfare concepts [2]. Both senior DoD leaders and PLA officials anticipate victory as ultimately belonging to the side with decision superiority through sensing and analyzing data more rapidly and accurately than their opponents. Data immediately tagged with a common data standard, cataloged, and securely stored within ABMS optimizes CCAs' potential in the HMT construct, giving commanders the lowest risk in the data architecture criterion.

### **Conditional**

The second dimension of MHC is the conditional dimension, defined by data understanding within its operating environment. Data understanding manifests as physical limitations embedded with CCA algorithms that constrain the timing, locality, and targeting of CEW operations. The success of a CEW CCA depends on its interpretation of the conditions of its operational environment, which will inevitably be mired in the fog of war.

Although the range of CCA's actions is bound by its algorithmic output, it must be capable of employing in environments lacking well-defined contexts. The critical question regarding how the CCA interprets its operating environment is, How can a machine be trained to suspect the



***The success of a CEW CCA depends on its interpretation of the conditions of its operational environment, which will inevitably be mired in the fog of war.***

truthfulness of its input and even its own training?

The concept of deception is antithetical to AI as a method of rapidly compiling and analyzing data. Data-saturated environments with true and false inputs must be considered when developing CCAs' data examination and processing cycles. Robust data understanding that detects deception is critical against AI-advanced adversaries like the PLA; standardized countermeasure EW tactics by CCAs will inevitably create opportunities for the PLA to use deception once discovered [19]. In other words, CEW CCAs must recognize when corrupted sensors introduced by the PLA are feeding them poisoned data.

One way to mitigate this vulnerability is through the CEW CCA's internalized AI adjudication to determine what deception is and is not through variable validity testing. However, standardizing these tests would provide opportunities for the PLA to deceive the test mechanisms.



Another medium-risk option to enhance the CCA's data understanding is by having committed, reachback EW experts, including engineers, EW officers, combat systems officers, and weapons systems officers, to analyze when, where, and how quickly the CEW CCA detected EW changes in the environment and whether its waveform countermeasures were successful. Deployed EW experts, as opposed to in-garrison, are more beneficial due to their proximity to the forward edge of the battlespace. This would allow them to deliver tactically relevant updates directly to the CCAs without relying on vulnerable datalinks. The closer human operators are to CCAs, the easier the logistics needed to adjust the CCAs' algorithms to act on the most accurate information, providing commanders the lowest risk in the data understanding criterion.

## Human-Machine Interaction

The third dimension of MHC is the human-machine interaction dimension, defined by data accessibility, visibility, and interoperability.

### Data Accessibility

Data accessibility entails the ability to access data for CCA mission execution. Data must be visible and accessible in a timely and relevant manner, at a minimum, to the authorizing commander. By having the authority to access the data driving the CCAs, the commander can assess the data's

reliability in driving CEW CCA systems. Data accessibility risk is lowered when data is accessible to both the theater commander employing the CCAs and the communities of interest (COIs) engaged in delivering the CCA data and assessing its data use in after-action studies. These COIs range from the in-garrison service intelligence agencies to the deployed EW professionals in theater. Data accessibility to all command echelons and stakeholders provides the lowest risk within the data availability criterion.

### Data Visibility

The second factor of the human-machine interaction dimension, data visibility through data interfaces, significantly influences the quality of HMT. Current demonstrations of pilot interactions with CCA prototypes leverage handheld tablets to send and receive operational data between the pilot and the machine [4]. Data that feeds CCA algorithms and manifests through CCA mission execution should ideally integrate with the Air Force's future ABMS infrastructure [20]. However, creating an effective user interface for ABMS data that feeds CCAs is a monumental challenge due to the massive amounts of sensor-to-shooter data expected to be collected [21].

Furthermore, as the amount of data-fueling CCA operations grows, the massive data influx to the human operator can lead to cognitive

overload, complacency, and loss of alertness [4]. Even though data interfaces with monitorable dashboards or a graphical user interface can be designed to alleviate this information overload, the traditional screen interfaces on tactical, 14-inch tablets and computers have physical restraints regarding their ability to display extensive real-time data to CCA human monitors.

One way to relieve humans from user interface limitations is by converging CCAs with neurotechnology, allowing bidirectional interaction between the human nervous system and the autonomous machine. Brain-computer interfaces (BCIs) would integrate the control of loyal wingmen into the human decision-making processes, accelerating their OODA loop and removing the task of designing CCA interfaces [4]. The U.S. Defense Advanced Research Projects Agency (DARPA) invests millions of dollars annually in BCI projects. DARPA's

“

***One way to relieve humans from user interface limitations is by converging CCAs with neurotechnology, allowing bidirectional interaction between the human nervous system and the autonomous machine.***

most recent noninvasive BCI program is its Next-Generation Nonsurgical Neurotechnology, which “aims to develop high-performance, bi-directional brain-machine interfaces for able-bodied service members” [22]. This interface enables technology to “control unmanned aerial vehicles and active cyber defense systems or teaming with computer systems to successfully multitask during complex military missions” [22]. The practical application of BCI research to CCA user interfaces provides the lowest risk of the data visibility criterion to authorizing commanders.

Data Interoperability

The third criterion to consider when evaluating a CCA’s human-machine interaction dimension is its data interoperability. A critical concern for effective HMT is the ability of the data driving the weapons system to be

interoperable within the larger system-of-systems context of ABMS. Data interoperability will ensure data flow and connectivity as ABMS evolves and expands, allowing data to fuel dynamic reassignment between humans and CCAs on the battlefield [7]. Data interoperability is enabled through common data standards used across not just the Air Force but also the other services, allies, and partners, with appropriately labeled releasability caveats on the data. All of these elements of data interoperability collectively provide commanders with the lowest risk of the data interoperability criterion.

DATA READINESS REPORTING TOOL

A summary of all seven data criteria and the respective standards that must be met to assign high, medium, or

low risk for CEW CCA data readiness is shown in Table 1. Commanders should use this tool to assess how much risk to mission they assume when employing AWS on the battlefield and understand how to best mitigate that risk with the proposed personnel, policy, and technology recommendations included in each risk level description. This tool provides ethical accountability of commander decisions by incorporating MHC over AWS in all the criterion assessments.

The proposed robotic wingmen readiness tool helps to ensure that CCA operations are not opaque to authorizing commanders. Rather than viewing CCA algorithm-driven behavior as a black box, data readiness ratings based on MHC dimensions position commanders to take responsibility for the AWS’s actions. The “low risk” data criteria column provides a strategic direction

Table 1. Proposed Robotic Wingmen Readiness Tool

COLLABORATIVE COMBAT AIRCRAFT DATA READINESS REPORTING TOOL					
MHC DIMENSION	HMT FACTORS	DATA CRITERIA	HIGH RISK	MEDIUM RISK	LOW RISK
Technological	Machine	Security	CDO data standards noncompliance	Public key cryptography	PQC standards
		Trust	CDAO metadata noncompliance	Limited metadata application	CDAO metadata compliance
		Architecture	Point-to-point data feeds	Networked command and control	AI/ML-based AMMS analytics
Conditional	Environment	Understanding	AI-validated algorithm	In-garrison expertise	Deployed expertise
Human-machine interaction	Human	Accessibility	Authorizing commander	Department of the Air Force data & EW COIs	DoD data COIs
		Visibility	Platform-specific user interface	Integrated ABMS user interface	Neurotechnology integration with ABMS data
		Interoperability	Minimal data standards met	Data standards limited to mission type	Contextualized to ABMS system of systems

for the DoD to optimize HMT through the ethical balance of autonomy and human interaction.

Although this article uses CEW CCAs as a vignette to explore the combat readiness of data, the seven proposed data criteria also apply to other robotic wingmen the DoD is fielding. The option for human control and verification is paramount in offensive autonomous weapons that can deliver lethal kinetic effects.

## CONCLUSIONS

Legacy readiness reporting criteria are insufficient to assess the combat readiness of the DoD's next generation of robotic wingmen. An appropriate framework to evaluate future robotic wingman readiness is through the three dimensions of MHC of autonomous systems, all of which characterize the data ultimately driving the AWS operations. Robotic wingmen assigned the lowest risk of each data criterion within the proposed CCA data readiness tool position CCAs to achieve optimal HMT with their human wingmen.

To realize the vision of this data readiness tool, future commanders and their staff will need to be data literate to accurately assess each data criterion. Data literacy and digital talent must no longer be siloed to specific Air Force Specialty Codes but rather foundational to all future

airmen. The Air Force must adopt a new readiness tool to reasonably hold commanders accountable for employing CCAs on the battlefield, and the DoD can use the Air Force's tool as its benchmark for assessing AWS across all services. As the DoD embraces AWS to meet the pacing challenge of China, service leaders must examine new ways to organize, train, and equip their members to best team with its future robotic wingmen. Like the 26th Secretary of the Air Force Frank Kendall stated in his 2022 Congressional hearing addressing Air and Space Force modernization efforts, "Change is hard, but losing is unacceptable" [23]. ■

## REFERENCES

- [1] U.S. DoD. "National Defense Strategy." Washington, DC: Office of the Secretary of Defense, p. 127, October 2022.
- [2] U.S. DoD. "Military and Security Developments Involving the People's Republic of China." 2023 *Annual Report to Congress*, Washington, DC: Office of the Secretary of Defense, pp. 40 and 95, 26 October 2022.
- [3] Bode, I., and H. Huelss. *Autonomous Weapons Systems and International Norms*. London, UK: McGill-Queen's University Press, 2022.
- [4] Rickli, J.-M. "Human-Machine Teaming in Artificial Intelligence-Driven Air Power: Future Challenges and Opportunities for the Air Force." *The Air Power Journal*, Fall 2022, vol. 8, [https://www.diaacc.ae/resources/2022\\_Jean\\_Marc\\_Rickli\\_Federico\\_Mantellassi\\_Human-Machine\\_Teaming\\_Air\\_Power.pdf](https://www.diaacc.ae/resources/2022_Jean_Marc_Rickli_Federico_Mantellassi_Human-Machine_Teaming_Air_Power.pdf), accessed on 12 March 2024.
- [5] United States Government Accountability Office (GAO). "Analysis of Maintenance Delays Needed to Improve Availability of Patriot Equipment for Training." GAO Report 18-447, Washington, DC, p. 7, June 2018.
- [6] Sanders, T. L., et al. *The Neurobiology of Trust*. Cambridge University Press, chapter 4, pp. 78 and 79, <https://www.cambridge.org/core/books/abs/neurobiology-of-trust/trust-and-human-factors/>

D51AD892F20EDA9404108BD66 49489E0#, 9 December 2021.

- [7] Madni, A. M., and C. C. Madni. "Architectural Framework for Exploring Adaptive Human-Machine Teaming Options in Simulated Dynamic Environments." *Systems* 6, no. 4, Spring 2018, pp. 3 and 15, <https://www.mdpi.com/2079-8954/6/4/44>, accessed on 5 January 2024.
- [8] Air Force Technology. "Collaborative Combat Aircraft." <https://www.airforce-technology.com/projects/collaborative-combat-aircraft-cca-usa/?cf-view&cf-closed>, accessed on 7 January 2024.
- [9] Vedula, P., et al. "Outsmarting Agile Adversaries in the Electromagnetic Spectrum." RAND Report A981-1, Santa Monica, CA: RAND Corporation, p. 5, 19 January 2023.
- [10] Clay, M. "To Rule the Invisible Battlefield: The Electromagnetic Spectrum and Chinese Military Power." *War on the Rocks*, <https://warontherocks.com/2021/01/to-rule-the-invisible-battlefield-the-electromagnetic-spectrum-and-chinese-military-power/>, 22 January 2021.
- [11] U.S. DoD. "DoD Data Strategy." Washington, DC: Office of the Secretary of Defense, pp. 1, 5, and 8, 30 September 2020.
- [12] Joint Chiefs of Staff. *Joint Publication (JP) 5-0. "Joint Planning,"* p. xiv, 1 December 2020.
- [13] National Cybersecurity Center of Excellence. "Migration to Post-Quantum Cryptography." NIST, <https://www.nccoe.nist.gov/sites/default/files/2023-08/mpqc-fact-sheet.pdf>, August 2023.
- [14] Stefanick, T. "The State of U.S.-China Quantum Data Security Competition." Brookings Institution, <https://www.brookings.edu/articles/the-state-of-u-s-china-quantum-data-security-competition/>, 18 September 2020.
- [15] Cybersecurity and Infrastructure Security Agency. "Quantum-Readiness: Migration to Post-Quantum Cryptography." Washington, DC, 21 August 2023.
- [16] U.S. DoD. "DoD Metadata Guidance, Version 1.0." Washington, DC: Chief Digital and Artificial Intelligence Officer, pp. 4 and 5, January 2023.
- [17] Hoehn, J. R. "Advanced Battle Management System." Congressional Research Service Report IF11866, vol. 5, 15 February 2022.
- [18] National Academies of Sciences. *Engineering, and Medicine, Advanced Battle Management System: Needs, Progress, Challenges, and Opportunities Facing the Department of the Air Force*. Washington, DC: The National Academies Press, p. 37, 2022.
- [19] Tangredi, S. J., and G. Galdorisi. *AI at War: How Big Data, Artificial Intelligence, and Machine Learning Are Changing Naval Warfare*. Annapolis, MD: Naval Institute Press, pp. 300, 301, 310, and 311, 2021.



[20] Johnson, T. R. "Emerging Tanker Roles and Risks in the Advanced Battle Management System Era." *Wild Blue Yonder*, <https://www.airuniversity.af.edu/Wild-Blue-Yonder/Article-Display/Article/2652095/emerging-tanker-roles-and-risks-in-the-advanced-battle-management-system-era/>, accessed on 15 January 2024.

[21] Wolfe, F. "Effective User Interfaces for ABMS a 'Momentous Challenge,' U.S. Space Force Says." *Defense Daily*, <https://www.defensedaily.com/effective-user-interfaces-abms-momentous-challenge-u-s-space-force-says/space/>, accessed on 14 October 2023.

[22] Willis, A. "Next-Generation Nonsurgical Neurotechnology." Defense Advanced Research Projects Agency, <https://www.darpa.mil/program/next-generation-nonsurgical-neurotechnology>, accessed on 16 February 2024.

[23] Secretary of the Air Force Public Affairs. "Kendall, Brown, Raymond Tell Congress \$194 Billion Budget Request Balances Risks, Quickens Transformation" *U.S. Air Force News*, <https://www.af.mil/News/Article-Display/Article/3012814/kendall-brown-raymond-tell-congress-194-billion-budget-request-balances-risks/>, accessed on 13 December 2023.

## BIOGRAPHIES

**CHRISTINA HAYHURST** is an active-duty U.S. Air Force (USAF) intelligence officer and instructor at the Squadron Officer School, with a follow-on assignment at the School of Advanced Air and Space Studies. Her operational assignments have provided intelligence support to Air Force bomber missions, Warfighter customers of the National Air and Space Intelligence Center, acquisition professionals within the Air Force Life Cycle Management Center, and the Air Force Special Tactics enterprise. Maj. Hayhurst holds a bachelor's degree in biochemistry from the USAF Academy, a master's degree in international security and economic policy from the University of Maryland, and a master's degree in military operational art and science from the Air Command and Staff College.

**CHRISTINE COVAS-SMITH** is the director of the Air Education and Training Command's Enterprise Learning Engineering (ELE) Center of Excellence at Joint Base San Antonio, Randolph Air Force Base, TX. She is leading the implementation of ELE as a sense-making framework for USAF development, including

increasing competency-based learning through systematic application of evidence-based principles, scientific methods, and practices from the learning sciences and education research and systems thinking to modernize learning. Dr. Covas-Smith holds a Ph.D. in applied psychology and cognitive action perception from Arizona State University.

**PATRICIA HARRIS** serves as the Writing Center Program lead at Air University and is a member of a core research group at the University of Bergen that focuses on the intersections of learning design, memory formation, and AI. She tested integrated perception ML models for Creative Synthetic and was a former professor, associate dean, technology director, and co-instructor for the Future Ideas and Weapons Research Task Forces at the Air War College. She wrote a book chapter focused on feedback strategies for Enlisted Professional Military Education that will be published by the U.S. Army Upgrade Program in 2025. Ms. Harris holds a master's degree in English literature and a Ph.D. in rhetoric and media studies.

# DSIAC WEBINAR SERIES

DSIAC hosts live online technical presentations featuring a DoD research and engineering topic within our technical focus areas. Visit our website to view our upcoming webinars.

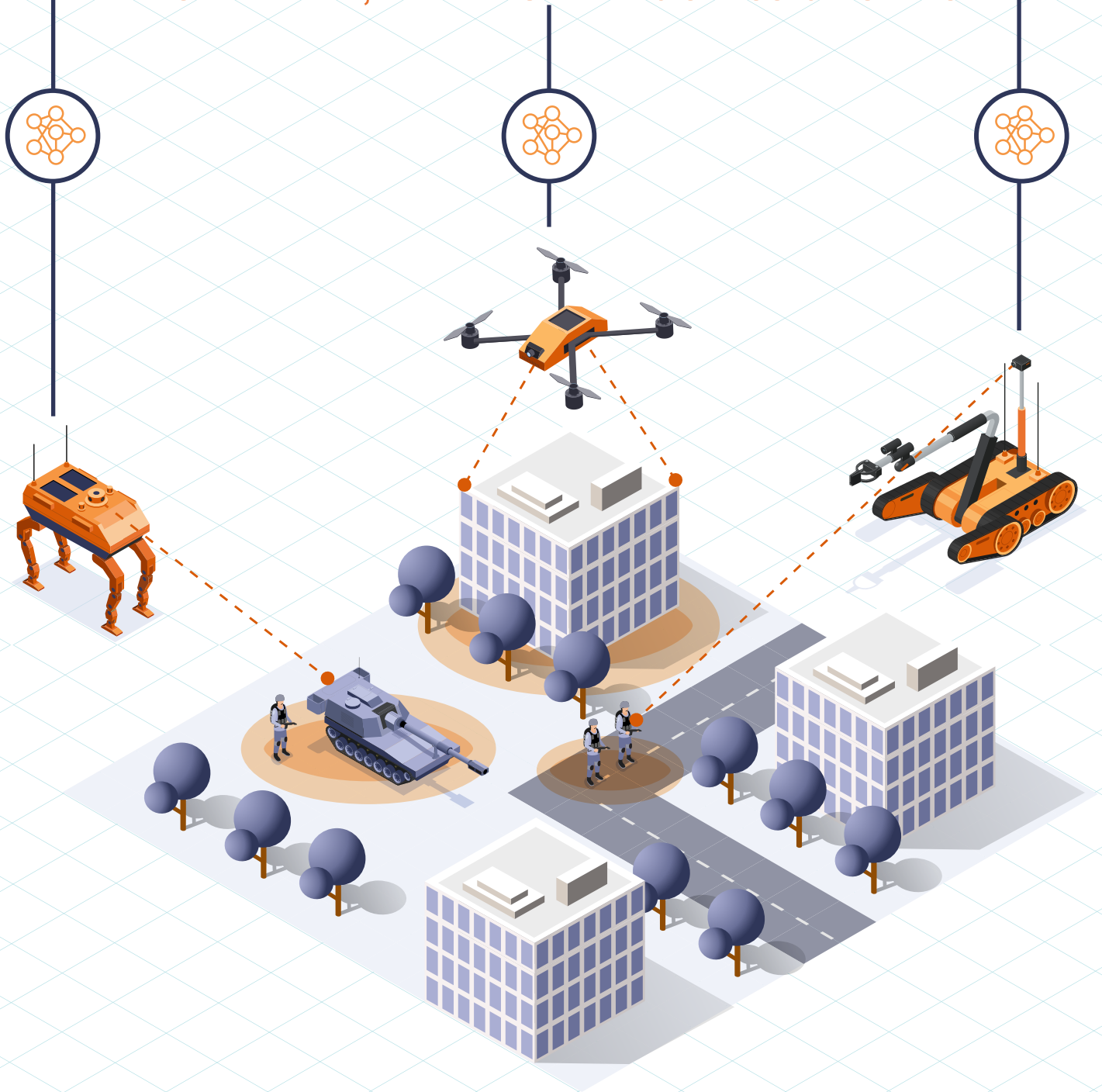
LEARN MORE



<https://dsiac.dtic.mil/webinars>

# MULTIAGENT FEDERATED LEARNING

## INTEROPERABILITY, AND VIRTUAL-PHYSICAL CO-SIMULATION



BY NIRMALYA ROY, JADE FREEMAN, MARK DENNISON, THERON TROUT, AND TIMOTHY GREGORY  
(PHOTO SOURCE: 123RF.COM)

## SUMMARY

The composition of different types of terrains and presence of a variety of objects and artifacts in a real environment are always evolving. Not every aspect of the environment can be learned, modeled, and synthesized a priori in an artificial intelligence (AI)-enabled, decision-making pipeline. If collaboration can be enabled among several deployed robots in different remote zones, it is possible to develop a generalized terrain and object detection model capturing greater uncertainty and variability in an environment. Conventionally, to realize this, data sharing between agents and servers is warranted; however, that may introduce a higher risk of an adversarial attack.

Federated learning can be a useful approach to mitigate this issue. The object detection model can be trained using federated learning in which training data will not be explicitly shared between the robots performing terrain reconnaissance in various geolocations. The robots can learn generalized features while training the model onboard and share the model updates from one location to another to support the collaborative training for distributed learning and real-time situational awareness. Also, simulating physical autonomous systems with virtual entities allows exploring complex interactions between collaborating agents at scale. Virtual-

physical co-simulation mitigates costly environments populated by large numbers of autonomous entities. A high-level overview of these approaches and a case study to overcome challenges of real-time adaptability and experiment scalability in multiagent teaming are presented in this article.

## INTRODUCTION

Current conflicts in Eastern Europe and the Middle East have demonstrated a new reliance on autonomous assets to augment various tasks in highly lethal battlefield environments. Therefore, it is imperative to understand how such robotic agents can efficiently and effectively communicate and collaborate among themselves as well as with human decision-makers to ensure battlefield dominance in real time. For example, the Ukrainian military has employed a hunter-killer style drone team where one unmanned aerial vehicle (UAV) is used to find enemy positions and another UAV drops a munition or the UAV itself is used as a disposable kinetic munition.

The research on multiagent teaming (MAT) is an area that remains less explored due to expensive resources to properly investigate on a realistic scale. Optimizing asset deployment in a vast area of interest (AOI) will require proper simulation in comparable scales and environments (e.g., open pastures and dense urban settings) or terrain

features (e.g., vegetation, forests, and deserts) to the real-world scenario. To that end, virtual reality can bridge the gap in scalability by augmenting real-world assets and physical environments in determining the resources required for desired effects.

“

***Current conflicts in Eastern Europe and the Middle East have demonstrated a new reliance on autonomous assets to augment various tasks in highly lethal battlefield environments.***

To achieve the maximum effect of MAT, it is imperative to understand how these collaborating autonomous agents can communicate with each other and complete tasks in an efficient manner. Emulating this scenario in the virtual world and combining it with the information obtained from various physical sources of intelligence can help to learn and understand the perception capabilities and collaborative behaviors of autonomous assets for mission success in the real world. Simulating physical autonomous systems with virtual entities allows exploring complex interactions between collaborating agents at scale. Furthermore, virtual-physical co-simulation mitigates costly



environments populated by large numbers of autonomous entities.

## COLLABORATIVE SITUATIONAL AWARENESS WITH MULTIAGENT FEDERATED LEARNING

Without sharing all the data, an object detection model learned in one environment (e.g., Site 1) can be leveraged to transfer to another remote location (e.g., Site 2). If collaboration can be enabled among several deployed robots in different zones, it is possible to develop a generalized terrain and object detection model that captures greater uncertainty and variability in an environment. Conventionally, to realize this, data sharing between agents and servers is warranted; however, this may introduce a security risk.

Federated class-incremental learning can be a useful approach to mitigate this issue. It allows multiple clients in a distributed environment to learn models collaboratively from evolving data streams where new classes arrive continually at each client. This technique helps accelerate the machine-learning (ML) model training process without directly sharing the raw data and pretrained models across multiple clients and only sharing the weighted average of the model parameters between them. This aids in collaborative training by sharing

knowledge from one geolocation to another while preserving privacy, minimizing the opportunities for data breaches, increasing the robustness of ML models, and reducing the training overhead and time substantially.

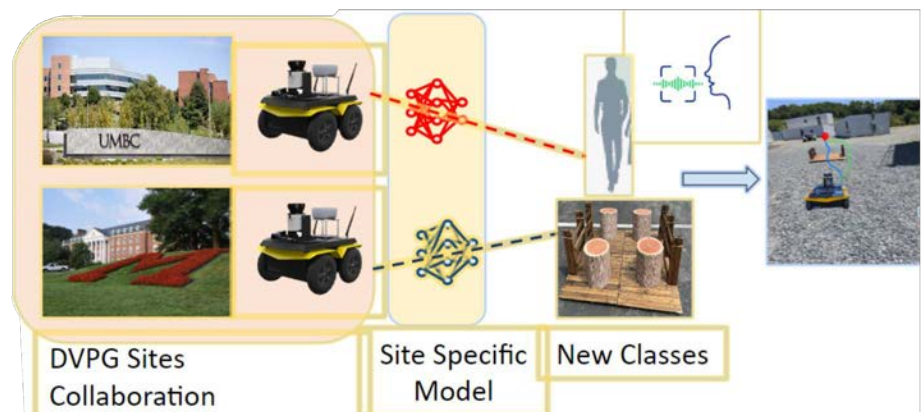
Investigating a remote zone reconnaissance scenario by relying on multiple distributed virtual remote sites using federated and continual learning is a novel research direction. The object detection model can be trained using federated learning in which training data will not be explicitly shared between robots performing terrain reconnaissance in various geolocations. The robots should learn generalized features while training the model onboard and share the model updates from one location to another to support the collaborative training for distributed situational awareness.

For example, an object such as a bridge seen at a location can be learned by the autonomous agents there and transferred with minimal

model parameters using federated learning to another location where the autonomous agents have not been trained to detect such a bridge, as shown in Figure 1. Collaborative training detects it in both locations by extending the learning securely to another autonomous agent in real time. This approach enables real-time collective situational awareness in an environment with minimal computing and training overhead, facilitating distributed and collaborative model training and remote learning.

## SEAMLESS INTEROPERABILITY AND SCALABILITY ACROSS HETEROGENEOUS ASSETS

Heterogeneity across robotic assets presents challenges for collaborative tasks. In addition to the inherent differences between unmanned ground vehicles (UGVs) and UAVs, unmanned X vehicles are manufactured by different vendors and have different



**Figure 1.** Secure Collaborative Training for ML Model Building (Source: N. Roy).

versions of autonomy stacks and software packages. For example, the simultaneous localization and mapping (SLAM) packages of UGV Jackals from Clearpath Robotics use Google's Robot Operating System (ROS) Cartographer; ROSbots from Husarion Inc. employ Hector SLAM packages; and modal AI RB5/Seeker drones rely on visual inertial odometry to support localization capabilities and volumetric pixels (voxels) for the mapping elements in their SLAM implementation. Interoperability across these packages while successfully transmitting the packets from one agent to another is a nontrivial research and development task.

A collaborative SLAM infrastructure involves data from multiple heterogeneous interoperable agents with different views of the same environment [1, 2]. The local views from multiple different agents are combined to create a global view of the sensing environment [3]. Moreover, calibrating and tuning SLAM parameters and sensitivity factors during the global map creation of an area are very important to successfully navigate a terrain using multiple robots. Therefore, software interoperability across ROS in virtual and physical environments is achieved, resulting in seamless connectivity and communication between multiple heterogeneous agents.

“

***Calibrating and tuning SLAM parameters and sensitivity factors during the global map creation of an area are very important to successfully navigate a terrain using multiple robots.***

## VIRTUAL-PHYSICAL CO-SIMULATION OF AUTONOMOUS NAVIGATION

Understanding the nature of terrain even before the deployment of autonomous robotic assets in the environment is important. Virtual-physical co-simulation allows optimal path planning to be determined in an AOI. Such a priori knowledge minimizes the risks and resources for allocating expensive autonomy assets in vast regions prior to deployment.

The first step in the simulation is to deploy all the autonomous agents in the virtual setting to simulate the path planning and coverage of sensing needed in the AOIs. The specific models of the UGVs/UASs in Gazebo and Unity can be deployed and the synthetic data generated from all the sensors, such as light detection and ranging (LiDAR); red, green, and blue;

and inertial measurement units for mapping and navigation.

Second, a mixed version of the prior setup is emulated with a few agents in the virtual environment covering a specific side of the terrain and some agents in the physical environment covering the other side. This designs the optimal path planning and minimizes the sensing overlap (and maximizes coverage) while detecting the objects, obstacles, adversaries, and other artifacts for making intelligent collaborative decisions using a swarm of UGVs and UASs. The interactions of robotic assets with virtual-physical environments are accurately modeled to represent behavior in the real world. The multiagent perception-action-communication loop cross-cutting between virtual and physical agents is being executed in the virtual-physical environment. Considering these factors, the robotic agent deployment planner can be designed while achieving the optimal path planning, sensing, and coverage of an area in the presence of adversaries.

An adaptive learning system within autonomous robotic agents is developed and integrated to enable real-time strategy adjustment in response to dynamic environmental changes and unforeseen challenges. With a focus on advancing the intelligence and adaptability of autonomous systems, a framework is created where UGVs/UASs can not only follow preplanned paths and

strategies but also learn from the environment and adjust their actions in real time. This adaptive learning is crucial in unpredictable environments where conditions can change rapidly or new obstacles and threats can emerge unexpectedly. Therefore, a key step is to develop a hierarchical objective-driven navigation system based on topological maps and novel learning algorithms to enable efficient path-finding and decision-making in environments with sparse rewards where feedback (rewards or penalties) for autonomous agents is minimal.

## REDUCING TOPIC DISSEMINATION OVERHEAD

In a virtual-physical co-simulation environment, information communication and exchange from physical and virtual space are established in multiagent scenarios for performing collaborative tasks such as object detection, scene perception, navigation, and route planning. Virtual and physical autonomous agents share the common world model as if they are collaborating in the real environment. In data sharing, it is imperative to reduce the overhead during ROS message passing and control with maximal information gain, minimal communication overhead, and maximum computing efficiency, as the communication network in the real world can be brittle and scarce.

One of the main challenges in the virtual-physical environment co-simulations is handling the underlying packet delays between simulators (Unity and Gazebo) and physical autonomous assets. Consider multiagent SLAM tasks where several autonomous agents explore and map an environment. Various ROS topics are being generated from each agent and transmitted over the wireless channels in the virtual and physical space as robots move around and scan an area to navigate. Even with a single agent, it is expensive to send all the ROS messages/topics from the robot's two-dimensional or three-dimensional (3-D) LiDAR, photographic imagery, or laser scanning to the master node. How can the agent send only essential information, metadata, or semantic knowledge of the environment? A solution to this question will reduce communication and computing overheads and reduce the network payload and delay in the simulation environment.

SLAM algorithms rely on flat representations of point cloud data and do not explore semantic relationships between objects, agents, structure, and their spatial arrangements [4]. To that end, a lightweight version of SLAM can be implemented, such as a spatial perception engine using dynamic scene graphs that capture the hierarchical relationships between the artifacts in the environment. This version can represent the high-level spatial concepts and relations rather than just lines, planes, points, and voxels.

Moreover, novel packet-filtering schemes and skip-window strategies can be employed to intelligently disseminate ROS topics from one agent to another, either situated in virtual or physical space, to reduce the number of messages being published and subscribed. This, in effect, will help improve the network's quality of service (QoS). An additional solution is leveraging the ROS2-based framework (masterless) in combination with topic aggregation and a selective unicast packet dissemination strategy instead of broadcasting all the ROS topics randomly from one agent to another or to a master node [1, 5].

“

***Novel packet-filtering schemes and skip-window strategies can be employed to intelligently disseminate ROS topics from one agent to another.***

## CASE STUDY

In this section, a case study performed as part of the August 2024 Summer Field Experiments at the U.S. Army Research Laboratory (ARL) Robotics Research Collaboration Campus (R2C2) in Graces Quarters (GQ), MD, will be discussed. AI-enabled decision making in multiple domains within complex and dynamic environments



will be addressed. Army capabilities like integrating data from all domains, reasoning across explicit and tacit knowledge, and supporting forces in both physical and information spaces will be covered.

The Army-relevant scenario of this field experiment involved a route reconnaissance, operational scenario representing challenges associated with autonomous navigation with seen and unseen object detection. It also involved avoidance in complex and noisy environments for which a commander conducted maneuver and intelligence decisions across multiple agents and modalities from different geolocations. The experiment showcased execution of remote voice command, obstacle avoidance, and a bridge-crossing task that was utilized to tie the specific elements of the research and contribute to aspects along the AI-enabled, intelligent decision-making cycle. The aspects of the following key research and development thrusts were addressed:

- Digital twin with photogrammetry rendering via Unity and synthetic data collection and annotation (Figure 2).
- Collaborative training with virtual-physical ML model building and minimal real data collection and annotation (Figure 2).
- Building privacy-aware new classes with learning from distant Distributed Virtual Proving Ground (DVPG) sites and federated class incremental learning.



**Figure 2.** Digital Twin for Virtual-Physical ML Model Building (Source: N. Roy).

- Autonomous navigation with object detection and avoidance and LiDAR semantic-segmentation-based navigation.
- Remote voice command with robot goal initialization using the DVPG network and voice enhancement with battleground noise.

As part of this collaborative research effort, components leveraging ARL's DVPG infrastructure to geographically distributed facilities and capabilities to perform joint mission and experiments across simulation and physical environments with robots were illustrated, as shown in Figure 2. The ability to interpret spoken instructions across the DVPG from a remote commander amidst noisy environmental conditions was also demonstrated.

In this collaborative remote robotics experiment, a UGV called Jackal was stationed at ARL's R2C2 in GQ's physical and virtual environments. Two other Jackals were stationed in two different geolocations—one in Maryland Robotics Center (MRC) at

the University of Maryland College Park (UMCP) and the other one in the Center for Real-Time Distributed Sensing and Autonomy (CARDS) at the University of Maryland Baltimore County (UMBC). Student commanders were present at UMCP and UMBC campuses to showcase how the remote learning, collaborative training, and automatic speech recognition (ASR) relying on multiple virtual and physical sites could help enhance the situational awareness in contested environments.

Three research contributions were the focus—ASR, virtual physical collaborative training, and new class learning via Federated Class Incremental Learning (FCIL). The virtual physical collaborative training was shown first, with ASR in a virtual GQ environment. UMCP student commanders were then asked to give the voice command to invoke the ARL ROS Unity Simulator at GQ.

A large set of photogrammetry data of the broken car and bridge in GQ was collected to integrate these objects

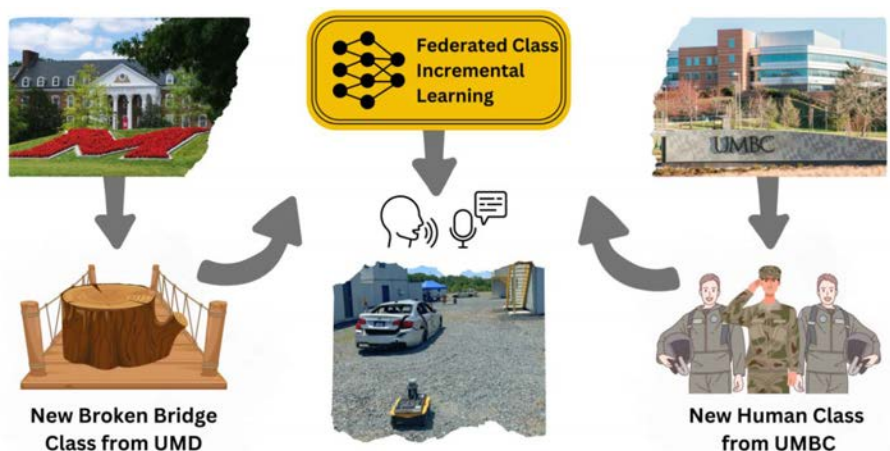
with the ARL's existing digital twin version of the GQ environment. Moreover, synthetic data about the broken car and bridge was collected. The LiDAR semantic segmentation in a virtual world was implemented to detect the broken car and bridge as well as autonomous navigation to avoid cars and successfully cross the bridge in a virtual GQ environment. The objective of the virtual experiment was to reinforce that meaningful synthetic data could be collected and annotated properly to implement collaborative training between virtual and physical sites. As part of the GQ virtual experiment, as shown in Figure 3, the rendering of the physical site of GQ inside Unity through the remote command and navigating the robot in physical (GQ) and virtual (Unity) spaces were executed by the student commander from UMCP using the DVPG network.

Next, the experiments showed autonomous navigation while avoiding the broken car and successfully crossing the bridge at the physical GQ site through collaborative training

and remote learning. An ROS board interface was implemented for publishing and subscribing all the ROS topics across multiple physical and virtual sites, as well as a domain adaptation technique between virtual and physical sites using synthetic and real datasets. As part of this experiment, student commanders at the UMBC DVPG site were asked to give the voice command to achieve the mission of bridge crossing at the physical GQ site while avoiding obstacles like a broken car. This case study represented the first contribution of the ASR and collaborative training

using real and synthetic data collected from multiple virtual and physical sites to improve situational awareness.

The second contribution was based on new class learning using the FCIL technique, as shown in Figure 4. In this part of the experiment, data was collected from the two remote physical sites of GQ DVPG—MRC at UMCP and CARDS at UMBC. Data of the new class bridge was also collected with tree logs at UMCP and another new class human at UMBC, as shown in Figure 5. (Note that the physical and virtual GQ agents were not



**Figure 4.** Building Privacy-Aware Machine Learning Model From Multiple Remote DVPG Sites (Source: N. Roy).



**Figure 3.** Remote Voice Command in a Virtual GQ Environment (Source: N. Roy).



**Figure 5.** Sharing Model Parameters for Unseen Classes From Remote DVPG Sites (Source: N. Roy).

trained with these two new classes during the first collaborative training experiment.)

Using the federated class incremental learning, the weighted average of the model parameters from UMCP and UMBC to the GQ physical site was transferred to see if the new class human and bridge with tree logs were being detected and avoided. This enabled the UGV Jackal in GQ to navigate autonomously—this time, the Jackal in GQ did not cross the bridge due to the fallen tree logs.

This case study depicted the second contribution. This field experiment attested the value of automatic speech recognition, collaborative training, and distributed remote learning from different geolocations in the presence of a digital twin, which has many potential applications for a multitude of civilian and military applications.

## RESEARCH AREAS OF OPPORTUNITY

Based on preliminary investigations presented here, there are many real-world challenges to explore in a virtual-physical co-simulation framework. The autonomous agents’ interoperability, calibration, and parameter tuning issues from software integration perspectives must be considered first. For path planning, navigation, and sensing with a minimal number of autonomous agents, the

optimal allocation of robotic assets is necessary where there are no unlimited autonomous assets available in the real world. Furthermore, universal data models that are compact, accessible across many programming languages, and efficient on the network need to be developed. Such models need to also support future scalability.

## CONCLUSIONS

Increased incorporation of autonomous-system interactions is anticipated. Autonomous drone swarms, unmanned ground vehicles, and packs of quadrupedal robots are no longer novel technologies. To ensure the use of these robots is effective, real-world challenges in the interactions among the collaborating robots need to be understood [6–9]. Research is being done on how detecting an object by an agent can be learned by other agents in another location by using federated learning to securely transfer with minimal model parameters.

“

*Autonomous drone swarms, unmanned ground vehicles, and packs of quadrupedal robots are no longer novel technologies.*

Further, software interoperability across heterogeneous robotic assets and real-time, 3-D semantic segmentation by transferring and sharing knowledge using federated learning models from distributed remote sites are being investigated. Finally, virtual-physical co-simulation is a novel experimentation approach that could be used toward scaling the number of agents to advance various robotics simulations in emerging application domains. ■

## ACKNOWLEDGMENTS

This research is supported by the U.S. Army Grant #W911NF2120076, Office of Naval Research Grant #N00014-23-1-2119, National Science Foundation (NSF) Research Experiences for Undergraduates (REU) Site Grant #2050999, and NSF Computer and Network Systems (CNS) EARly-concept Grants for Exploratory Research (EAGER) Grant #2233879.

The authors would like to acknowledge the following contributions:

- ARL researchers—Dr. Niranjana Suri, Dr. Adrienne Raglin, Dr. Carl Busart, Dr. Stephanie Lukin, Dr. Claire Bonial, and Dr. Felix Gervits.
- UMBC and UMCP faculty—Dr. Anuradha Ravi, Dr. Abu Zaher Faridee, Dr. Zahid Hasan, Dr. Frank Ferraro, Dr. Cynthia Matuszek, Dr.



Aryya Gangopadhyaya, Dr. Derek Paley, and Dr. Carol Espy Wilson.

- UMBC and UMCP students—  
Masud Ahmed, Emon Dey, Jumman Hossain, Saeid Anwar, Indrajeet Ghosh, Gaurav Shinde, Snehalraj Chugh, Sadman Sakib, Vinay Krishna Kumar, and Shubham Ojha.
- Undergraduate student researchers as part of the NSF REU program—  
Hersch Nathan and Adam Goldstein.

## REFERENCES

- [1] Dey, E., J. Hossain, N. Roy, and C. Busart. "SynchroSim: An Integrated Co-Simulation Middleware for Heterogeneous Multi-Robot System." The 18th International Conference on Distributed Computing in Sensor Systems, 2022.
- [2] Mohammad Saeid Anwar, M. S., A. Ravi, E. Dey, G. Shinde, I. Ghosh, J. Freeman, C. Busart, A. Harrison, and N. Roy. "CoOpTex: Multimodal Cooperative Perception and Task Execution in Time-critical Distributed Autonomous Systems." The 21st International Conference on Distributed Computing in Smart Systems and the Internet of Things, 2025.
- [3] Anwar, M. S., E. Dey, M. K. Devnath, I. Ghosh, N. Khan, J. Freeman, T. Gregory, N. Suri, K. Jayarajah, S. R. Ramamurthy, and N. Roy. "HeteroEdge: Addressing Asymmetry in Heterogeneous Collaborative Autonomous Systems." The IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems, pp. 575–583, 2023.
- [4] Ahmed, M., Z. Hasan, A. Z. M. Faridee, M. S. Anwar, K. Jayarajah, S. Purushotham, S. You, and N. Roy. "ARSFineTune: On-the-Fly Tuning of Vision Models for Unmanned Ground Vehicles." The 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things, IEEE, 2024.
- [5] Dey, E., M. Walczak, M. S. Anwar, N. Roy, J. Freeman, T. Gregory, N. Suri, and C. Busart. "A Novel ROS2 QOS Policy-Enabled Synchronizing Middleware for Co-Simulation of Heterogeneous Multi-Robot Systems." The 32nd International Conference on Computer Communications and Networks, pp. 1–10, 2023.
- [6] Hossain, J., A. Z. Faridee, D. Asher, J. Freeman, T. Trout, T. Gregory, and N. Roy. "QuasiNav: Asymmetric Cost-Aware Navigation Planning With Constrained Quasimetric Reinforcement Learning." The IEEE International Conference on Robotics and Automation (ICRA), 2025.

[7] Shinde, G., A. Ravi, E. Dey, J. Lewis, and N. Roy. "TAVIC-DAS: Task and Channel-Aware Variable-Rate Image Compression for Distributed Autonomous System." The 4th IEEE Workshop on Pervasive and Resource-constrained Artificial Intelligence, colocated with the 23rd IEEE International Conference on Pervasive Computing and Communications (PerCom), 2025.

[8] Dey, E., A. Ravi, J. Lewis, V. Kumar, J. Freeman, T. Gregory, N. Suri, C. Busart, and N. Roy. "DACC-Comm: DNN-Powered Adaptive Compression and Flow Control for Robust Communication in Network-Constrained Environment." The 17th International Conference on Communication Systems & NETWORKS (COMSNETS), 2025.

[9] Hossain, J., and N. Roy. "Learning Optimal Policies With Quasi-Potential Functions for Asymmetric Traversal." The 42nd International Conference on Machine Learning (ICML'25), 2025.

## BIOGRAPHIES

**NIRMALYA ROY** is a professor in the Information Systems Department and director of the Mobile, Pervasive, and Sensor Computing Lab at UMBC; associate director of CARDS at UMBC; and a co-PI on an ArtIMAS (AI and Autonomy for Multi-Agent Systems) cooperative research agreement from ARL in collaboration with the UMCP. His current research interests include use-inspired AI/ML and human-centric data science with applications to smart health, cyber-physical systems, Internet of Things, robotics, and autonomy. Prior to joining UMBC, he was a clinical assistant professor at Washington State University, a research staff member at the Institute for Infocomm Research in Singapore, and a postdoctoral fellow at the University of Texas at Austin. Dr. Roy holds a B.E. in computer science and engineering from Jadavpur University, India, and an M.S. and Ph.D. in computer science and engineering from the University of Texas at Arlington.

**JADE FREEMAN** is the chief of the Battlefield Information Systems Branch at the U.S. Army Combat Capabilities Development Command (DEVCOM), ARL, where she manages research projects on cross-reality technology, large language model, computer vision, and AI resilience and assurance cases. Dr. Freeman holds a Ph.D. in statistics from George Washington University.

**MARK DENNISON** is the information dynamics team lead for the Battlefield Information Systems Branch at DEVCOM, ARL, where he explores how augmented and mixed reality technologies can enhance information visualization and interaction, sharing, and exploitation to enable shared situational understanding across highly decentralized maneuver forces and allow warfighters to collaborate more effectively with robotic systems. Previously, he studied how heterogeneous physiological sensor information can be used to predict motion sickness onset and severity in head-mounted display users.

Dr. Dennison holds a bachelor's degree in psychology, a master's degree in cognitive neuroscience, and a Ph.D. in psychology from the University of California, Irvine.

**THERON TROUT** is vice president and chief operating officer of Stormfish Scientific Corporation, where he focuses on enabling large-scale, distributed environments interconnecting users in virtual-, augmented-, and mixed-reality technologies to perform military-relevant analysis and decision-making activities. Mr. Trout holds a B.S. in computer science, with minors in physics and mathematics, from Marshall University.

**TIMOTHY GREGORY** is an electronics engineer in the Battlefield Information Systems Branch at DEVCOM, ARL, where he works on various software and hardware projects in robotics, unattended ground sensors, database systems, geographic information systems, communications protocols, and sensor simulation and network systems. He also oversees and manages research experiments and field tests at Army and multinational coalition military exercises. Mr. Gregory holds a B.S. in computer science from UMCP.

## SHARE YOUR EXPERTISE

If you are a contributing member of the defense systems community and are willing to share your expertise, you are a DSIAC subject matter expert.

<https://dsiac.dtic.mil/subject-matter-experts>

# Discover the **value** of sharing your **DoD-funded research...**



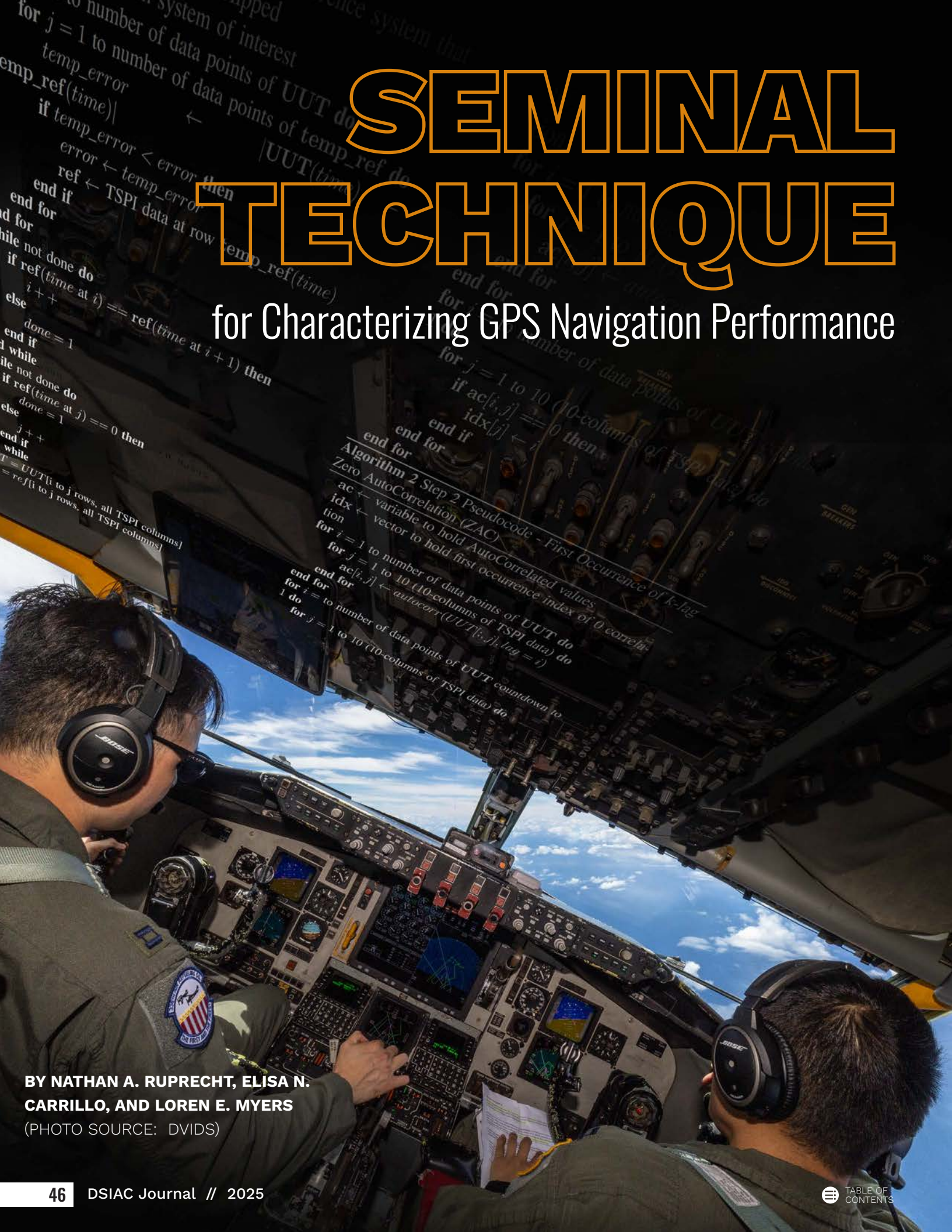
**R&E GATEWAY**

POWERED BY **DTIC**

<https://submit.dtic.mil/submit>

Defense Technical Information Center (DTIC) | Fort Belvoir, VA





# SEMINAL TECHNIQUE

for Characterizing GPS Navigation Performance

BY NATHAN A. RUPRECHT, ELISA N.  
CARRILLO, AND LOREN E. MYERS  
(PHOTO SOURCE: DVIDS)



## SUMMARY

**T**his article presents a standardized technique to evaluate the navigation performance (position, velocity, and principal axes) of an airborne system and develop an efficient methodology to define the number of target data points to process during the planning phases of testing. This technique gave statistical credibility and rigor to the test while saving time and funds and with the minimum amount of data to achieve significant results.

Three units under test (UUTs) were used that underwent ground and flight tests alongside a reference system. The proposed methodology involved calculating the Pearson correlation coefficient to find intervals of random error. These intervals were used to pick out the data points to use for calculating error. For a more complete analysis, the error calculated using this methodology was shown alongside available data.

The difference between the two methods of calculating error was nearly negligible across all UUTs and time, space, and position information (TSPI) columns of interest, with multiple instances of 0% difference.

## INTRODUCTION

Since its inception, engineers and scientists have continuously

worked to enhance the accuracy and precision of global positioning system (GPS) and inertial navigation system (INS) technologies [1]. These efforts have focused on overcoming inherent system limitations and environmental challenges, enabling reliable performance in a wide range of applications. GPS and INS systems, while individually robust, are often integrated into hybrid configurations to capitalize on their complementary strengths. For example, embedded GPS and INS packages combine the long-term accuracy of GPS with the short-term stability of INS, allowing for multiple navigation modes and improved system redundancy [2–5]. This integration mitigates errors associated with each standalone system, such as GPS signal degradation in obstructed environments or INS drift over extended durations.

To further enhance performance, many navigation systems incorporate real-time error estimation algorithms. These algorithms leverage statistical models to quantify and correct for errors in position, velocity, and orientation during operation [6, 7]. By continuously refining their estimates, such systems provide users with more accurate and reliable navigation solutions. For instance, state-of-the-art error estimation techniques often utilize Kalman filtering or similar probabilistic approaches to fuse sensor data and predict system accuracy [8, 9].

“

***To further enhance performance, many navigation systems incorporate real-time error estimation algorithms.***

Accuracy measurements for navigation systems are typically evaluated in the Earth-centered, Earth-fixed (ECEF) coordinate system, which defines errors in the  $X$ ,  $Y$ , and  $Z$  directions as  $\Delta X_i$ ,  $\Delta Y_i$ , and  $\Delta Z_i$ , respectively. These errors can result from various sources, including signal multipath, atmospheric disturbances, and sensor noise. Quantifying these errors is critical for system validation, particularly in applications with stringent performance requirements, such as aviation and autonomous vehicle navigation. While numerous metrics exist for assessing navigation accuracy, two are commonly employed—spherical error probable (SEP) and root mean squared (RMS) error.

SEP provides a probabilistic measure of positional accuracy, defining the radius within which a certain percentage of positional estimates fall [10]. However, this article focuses on RMS, a metric that directly quantifies the average magnitude of errors in all three spatial dimensions. RMS offers a straightforward and comprehensive means of comparing system performance under various conditions.

When designing a test to evaluate GPS accuracy, a central question arises: “How much data is enough?” This question often sparks debate among stakeholders. Engineers and system developers, aiming to maximize statistical reliability, typically advocate for collecting as much data as possible. Conversely, financial analysts and project managers, focused on minimizing costs and resources, push for constraints on data collection efforts. Striking a balance between these competing priorities requires careful statistical planning during the test design phase. By incorporating foresight into the expected number of observations, it is possible to achieve a rigorous analysis without excessive resource expenditure.

One of the foundational statistical tools for addressing sample size requirements is the Student’s t-Distribution, a widely used approach for analyzing small datasets. This is particularly effective when the goal is to estimate the mean of a normally

distributed population where the sample size is small and the population standard deviation is unknown [11]. A commonly cited rule of thumb suggests that a minimum of 30 data points is sufficient to utilize the t-Distribution effectively [12]. This threshold is rooted in the central limit theorem, which states that the sampling distribution of the mean approaches normality as the sample size increases, even if the underlying population distribution is not perfectly normal [13, 14]. For small samples, the t-Distribution provides a robust framework with well-defined properties, including  $n - 1$  degrees of freedom and an expected mean of zero, variance of 1, and standard normal behavior, denoted as  $N \sim (0,1)$ .

However, determining the appropriate sample size involves more than adhering to arbitrary thresholds. The choice of 30 data points assumes that the data represent the population and that the chosen distribution accurately models the underlying behavior. If these assumptions are violated—such as when the data exhibit significant skewness, kurtosis, or outliers—additional considerations must be made [15, 16]. For example, heavily skewed distributions may require larger sample sizes to achieve reliable results, while data with outliers may necessitate robust statistical methods or preprocessing to mitigate their influence [17].

An equally critical aspect of sample size determination is defining what constitutes a valid data point. In GPS accuracy testing, a “data point” often corresponds to a discrete event or observation, such as a position fix or navigation update. The temporal and spatial resolution of these observations can significantly impact the test results. For instance, higher-frequency data collection may capture transient anomalies that lower-frequency sampling would miss, while excessive sampling may introduce redundancy without adding meaningful information [18]. Understanding these trade-offs is essential for ensuring that the collected data is sufficient and efficient for the intended analysis.

While many test designs rely on post-hoc statistical power analysis to validate results after data collection, this approach can be inefficient and costly. By integrating statistical planning into the test design process, planners can optimize data collection strategies, reduce resource consumption, and improve the overall reliability of their findings. Techniques such as power analysis, sensitivity analysis, and simulation-based methods can help refine sample size estimates and identify the minimum data requirements for achieving statistically significant results [19, 20].

Ultimately, the process of determining how much data is enough depends on a nuanced understanding of the statistical properties of the chosen

“

***One of the foundational statistical tools for addressing sample size requirements is the Student’s t-Distribution, a widely used approach for analyzing small datasets.***

methodology and the specific objectives of the test. This article aims to address these challenges by proposing a systematic approach to defining valid data points and evaluating sample size requirements in the context of GPS accuracy testing.

With a t-Distribution and the widely accepted target of 30 data points established, the next step is to clearly define what constitutes a valid sample. While the conceptual process of selecting data points may appear straightforward, the practical implementation often introduces complexities that can compromise the validity and reliability of results [21]. A data sample in GPS accuracy testing is typically tied to a specific event like a position update or navigation fix recorded during the operation of the system under test. However, the characteristics of these events—such as their temporal spacing, environmental conditions, or measurement noise—can significantly influence the outcomes of subsequent analyses.

A key challenge in this process is ensuring that the selected data points are representative of the underlying system behavior and relevant to the test objectives. For instance, if the test environment includes a mix of benign and degraded GPS conditions, the sampling strategy must account for this variability to avoid skewed results. Oversampling events in benign conditions could mask the system's true limitations, while focusing disproportionately

on degraded scenarios might inflate error metrics and lead to overly conservative conclusions. Additionally, the temporal distribution of data points plays a crucial role. Sampling intervals that are too short may introduce autocorrelation effects, where consecutive measurements are highly dependent, violating statistical independence assumptions [22]. Conversely, overly long intervals may fail to capture transient system behaviors critical to navigation performance evaluations.

“

***If the test environment includes a mix of benign and degraded GPS conditions, the sampling strategy must account for this variability to avoid skewed results.***

In practice, many test planners adopt a comprehensive approach, utilizing all available data to calculate system accuracy. This approach often ensures that the results reflect the full spectrum of operating conditions encountered during the test. However, it can also lead to inefficiencies, such as processing redundant or irrelevant data, and may obscure critical insights into system performance under specific conditions. Moreover, relying on post-execution statistical power assessments, as is often the case, limits the ability to

adapt data collection strategies in real-time, potentially compromising test outcomes [23].

To address these challenges, it is necessary to develop a systematic methodology for identifying valid data points. Such a methodology should consider the statistical properties of the data and the practical constraints of the test environment. For example, criteria for data selection might include thresholds for measurement uncertainty, filtering for specific environmental conditions, or stratification by navigation mode or operational phase. By incorporating these criteria into the test design process, planners can ensure that the selected data points provide meaningful insights while maintaining statistical rigor.

The assumption that 30 data points are sufficient for statistical significance also warrants scrutiny. Although this threshold is often a general rule of thumb, its applicability depends on several factors, including the underlying distribution of the data, the presence of outliers, and the desired level of confidence in the results [24–26]. For example, in cases where the data exhibit significant skewness or heavy tails, larger sample sizes may be required to achieve reliable estimates of central tendency and dispersion. Conversely, in well-controlled environments with low measurement noise, fewer data points may suffice to achieve the same level of statistical power [27].



The equations and concepts that underlie GPS and INS performance evaluations have been extensively applied in ground [28, 29] and airborne [30] systems, serving as the basis for establishing their suitability as reference systems. These efforts have culminated in developing highly accurate and well-characterized navigation systems, many of which remain benchmarks in the field [31].

However, a critical question often overlooked in these applications is what methodology is used to select the data points that form the basis of performance evaluations. While it is common practice to calculate the three-dimensional (3-D) error using all available data, this approach, while comprehensive, is not without drawbacks. Processing large datasets indiscriminately can lead to inefficiencies in terms of computational resources and test planning efforts, particularly when cost and time constraints are significant.

In specialized fields like navigation performance testing, where technical applications are highly specific, much of the expertise and methodologies is often passed down informally through experience rather than systematically documented. This reliance on institutional knowledge can create challenges in maintaining consistency and rigor, particularly when personnel turnover results in the loss of statistical analysis subject matter experts. As



***Processing large datasets indiscriminately can lead to inefficiencies in terms of computational resources and test planning efforts, particularly when cost and time constraints are significant.***

such, the development of a standardized, repeatable methodology for test planning is not just beneficial but essential for sustaining high standards within the testing community.

The traditional approach of leveraging all available data for analysis undoubtedly ensures that performance metrics represent the test conditions. However, this exhaustive approach can obscure opportunities to optimize the testing process, especially during the planning phase. A more efficient strategy involves defining what constitutes a valid data point in the context of the test objectives and using this definition to estimate the minimum sample size required for statistically significant results. By narrowing the scope to a subset of data that is representative and relevant, test planners can achieve dual goals of maintaining statistical rigor and reducing unnecessary resource expenditures.

This article proposes a novel methodology aimed at addressing these challenges, particularly by moving beyond the traditional reliance on Student's t-Distribution assumptions. While the t-Distribution provides a robust framework for estimating population parameters under certain conditions, its utility in practical test planning is limited when the criteria for data selection and the characteristics of the dataset are poorly defined. By incorporating a systematic process for identifying valid data points, grounded in statistical tools like error interval calculations and correlation analyses, this methodology enables test planners to answer the critical question of how much data is enough with greater precision.

The remainder of this article details the development and application of this methodology, illustrating its potential to streamline the planning phase of navigation performance tests. The proposed approach offers a way to predefine data collection requirements that ensures efficiency and reliability, ultimately supporting the broader objective of delivering rigorous and cost-effective evaluations of GPS and INS systems.

## METHODOLOGY

This section depicts the bulk purpose and focus of this article by presenting the approach and technique principals that were applied to flight test data.

Data Characteristics

TSPI was collected at the default data rates for each component: 1 Hz for GPS and 10 Hz for INS. These data sets were derived from multiple UUTs, manufactured by different vendors, and deployed on a fixed-wing, turboprop transport aircraft. The aircraft, modified for military testing, could carry cargo and personnel while supporting research and development evaluations of the onboard navigation systems. (Specifics beyond this are intentionally omitted for security reasons and to keep the emphasis of this article on the proposed process.)

The characteristics of the tests performed on each UUT in ground-based and airborne environments under varying conditions are shown in Table 1 and include the hours of ground test/flight tests, benign or degraded environments, range of velocity and position, and flight time in a degraded environment for 85 hours of testing across three different UUTs.

Each TSPI dataset consisted of 10 distinct columns:  $V_x$ ,  $V_y$ ,  $V_z$  (velocity components),  $P_x$ ,  $P_y$ ,  $P_z$  (ECEF position components), roll, pitch, yaw, and time. Here,  $P$  represents the ECEF positions along the three Cartesian axes, while yaw describes the rotation of the body axis. The datasets were intentionally standardized across UUTs to maintain consistency in comparisons.

The Ultra-High Accuracy Reference System (UHARS), built and maintained by the 746th Test Squadron (746 TS) at Holloman Air Force Base (AFB), NM, served as the standard reference system for these evaluations. UHARS is regarded for its ability to deliver highly precise positional and navigational data, even in challenging environments, making it an ideal benchmark for assessing the performance of emerging navigation systems. A critical characteristic of a reference system is its inherent accuracy and precision relative to the UUT. For robust testing, it is generally accepted that the reference

system’s accuracy should exceed that of the UUT by at least an order of magnitude. This ensures that the reference system’s contribution to overall error is negligible when quantifying the performance of the UUT.

The technical justification for this rule of thumb is rooted in the propagation of uncertainty during error calculations. If the reference system’s error approaches the error of the UUT, it becomes challenging to differentiate between the inherent inaccuracies of the UUT and the limitations of the reference system itself [32, 33]. By

“  
*UHARS is regarded for its ability to deliver highly precise positional and navigational data, making it an ideal benchmark for assessing the performance of emerging navigation systems.*

Table 1. Demographics of Flight Test Data From Three UUTs Flown on the Same Aircraft Across Multiple Sorties (Source: N. Ruprecht)

DEMOGRAPHIC	CONFIGURATION	UUT1	UUT2	UUT3	TOTAL TEST TIME (HR)
Static vs. dynamic	Time in flight (hr)	9.4	30.3	18.81	58.51
	Time on ground (hr)	5.79	15.87	5.06	26.71
Environment	Time with degraded GPS (hr)	6.38	17.47	11.06	34.91
	Time in clear air (hr)	8.8	28.7	12.81	50.31
Other characteristics	Flight time with degraded GPS (hr)	5	15.96	11.06	32.01
	Range of velocity (m/s)	117.65	124.33	124.33	
	Range of 3-D position (m)	3163.79	3163.79	2963.32	
	Total UUT time (hr)	15.18	46.17	23.87	85.22

maintaining an accuracy margin of 10:1, the reference system effectively isolates the UUT's performance characteristics.

For instance, if the reference system's positional accuracy is  $\sim 1$  cm, it can reliably assess a UUT with positional errors near 10 cm or greater without introducing significant ambiguity into the results. The referenced "10 $\times$  accuracy rule" is also rooted in practical testing heuristics, where the reference system is expected to be an order of magnitude more accurate than the system under test. This is also known as the Gagemaker's Rule or Rule of Ten. It has been associated with military standards like MIL-STD-120 released in 1950 but has shifted to a 25% tolerance and is becoming increasingly challenging to maintain in all cases [34].

Additionally, ISO/IEC 17025, a global standard for testing and calibration laboratories, highlights the need for traceability and rigorous uncertainty analysis to ensure that the reference system's errors do not compromise the integrity of the test results [35]. By maintaining this high level of accuracy, UHARS ensures that observed deviations during testing can be confidently attributed to the performance of the UUTs rather than errors introduced by the reference system. This approach upholds the statistical credibility and reliability required for evaluating navigation system performance, particularly

in environments where precision is critical.

## Proposed Technique

There are multiple sources of errors related to GPS. These errors are made up of time-correlated and nontime-correlated components [36, 37]. Inertial systems also have correlations due to their inherent growth in inertial sensor error [38]. To characterize a system, the errors that are evaluated at a certain time should be uncorrelated (correlation coefficient zero) of any other [39, 40]. Note that this process requires the data to have zero correlation but not zero dependence, where correlation is the measure of linear dependence. Since the UUT and reference system will have similar distribution of errors for a given column and one does not affect or impact the other, the systems are assumed independent and identically distributed.

The overall process presented to evaluate the navigation performance of airborne systems follows four steps:

1. Align, interpolate, and clip data so that data analyzed records the UUT and reference system.
2. Find correlation coefficient for  $k$  number of lags and save the first occurrence  $k$  of zero autocorrelation (ZAC) for each column of interest.
3. Calculate RMS error at modulus intervals of previous step between

the UUT and corresponding reference data point.

4. If the target system specification (population mean) is known, use (a). Otherwise, use (b), where (a) runs a hypothesis test given these sample errors, target specification, standard deviation, and number of samples collected at those intervals (thus saving and presenting the hypothesis test result and corresponding t-statistic and one-sided p-value for each column of interest) and (b) creates a single tailed, upper bound confidence interval for these sample errors, number of samples collected, and chosen  $\alpha$  for each column of interest. Also calculated is the RMS error (Step 3) using all available data to show side-by-side comparison of techniques and that the two are comparable.

Algorithm 1 (Figure 1) is used to preprocess the TSPI data collected during each test. If the UUT and reference system are not precisely synchronized for recording and sample at different GPS times, this script aligns the two sets of data by finding the minimum error between the time columns and assigning that row to the final reference variable to be used. If the sampling frequencies of the two systems are different, the script still accounts for this using minimum error and may then list the same row multiple times. The "while" loop is used in case the system recording starts or ends before/after



---

**Algorithm 1** Step 1 Pseudocode - Align, Extrapolate, and Clip

---

```
temp_ref ← TSPI from reference system
ref ← final variable of TSPI from reference system that
is aligned, extrapolated, and clipped
UUT ← TSPI from system of interest
for i = 1 to number of data points of UUT do
    for j = 1 to number of data points of temp_ref do
        temp_error ← |UUT(time) -
temp_ref(time)|
        if temp_error < error then
            error ← temp_error
            ref ← TSPI data at row temp_ref(time)
        end if
    end for
end for
while not done do
    if ref(time at i) == ref(time at i + 1) then
        i ++
    else
        done = 1
    end if
end while
while not done do
    if ref(time at j) == 0 then
        done = 1
    else
        j ++
    end if
end while
UUT = UUT[i to j rows, all TSPI columns]
ref = ref[i to j rows, all TSPI columns]
```

---

**Figure 1.** Algorithm 1 (Source: N. Ruprecht).

the other system. It will log the index for the first time the two variables show alignment and the last, therefore clipping the **UUT** and **ref** variables.

After preprocessing and data cleaning, each column of interest is compared to itself to check for correlation (autocorrelation). Each index is compared to every other as the data is shifted by its entire length to find the Pearson correlation coefficient for each lag. Given measurements  $Y_1, Y_2, \dots, Y_N$  at time  $X_1, X_2, \dots, X_N$ , the lag  $k$  autocorrelation function is defined in Equation 1 as follows:

$$r_k = \frac{\sum_{i=1}^{N-k} (Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^N (Y_i - \bar{Y})^2}. \quad (1)$$

Algorithm 2 (Figure 2) finds the coefficient of each data sample compared to a shifted vector (lag) of itself using a Python function that utilized this equation. There may be multiple instances of zero correlation between data points and therefore multiple intervals that can be used to build a data array. To have the least amount of time between points, this script looks for the first occurrence of zero correlation for each TSPI column.

After the smallest interval between uncorrelated data, Algorithm 3 (Figure 3) goes back through the UUT data to calculate error. Using the modulo function, if  $i$  is a multiple of the index for a given column, that data point in UUT is used against the reference system at that same point in time (or closest to referring to Algorithm 1 in aligning). With arrays of absolute errors, the dataset can now be analyzed knowing its errors are random or uncorrelated. In this case, the RMS, standard deviation, and number of data points of each column ( $N$ ) are saved to **Stats**.

Finally, Algorithm 4 (Figure 4) shows the final piece of characterizing the UUT.

Using the **Stats** variable, a single-mean, t-statistic is calculated based on error as defined by Equation 2, where  $x$  is the sample mean error given by the column's RMS,  $s$  is the sample standard deviation,  $N$  is the number of samples used, and  $\mu$  is the population mean error given by the target specification RMS for each column:

$$t = \frac{x - \mu}{s/\sqrt{N}}. \quad (2)$$

With a t-statistic, the single-sided p-value can be determined using a Python function in the statistics library. With these two knowns, a hypothesis test is run with the null and alternative defined as  $H_0: x \geq \mu$  and  $H_a: x < \mu$ , with rejecting  $H_0$  if  $p < \alpha$  and  $t < 0$ , where  $\alpha = 0.05$  in this case. The reason for this order is

---

**Algorithm 2** Step 2 Pseudocode - First Occurrence of k-lag Zero AutoCorrelation (ZAC)

---

```
ac ← variable to hold AutoCorrelated values
idx ← vector to hold first occurrence index of 0 correlation
for i = 1 to number of data points of UUT do
  for j = 1 to 10 (10-columns of TSPI data) do
    ac[i, j] ← autocorr(UUT[:, j], lag = i)
  end for
end for
for i = to number of data points of UUT countdown to 1 do
  for j = 1 to 10 (10-columns of TSPI data) do
    if ac[i, j] == 0 then
      idx[j] ← i
    end if
  end for
end for
```

---

**Figure 2.** Algorithm 2 (Source: N. Ruprecht).

---

**Algorithm 3** Step 3 Pseudocode - Calculate Descriptive Statistics at Index Intervals

---

```
Stats ← variable to hold RMS, standard deviation of RMS, and number of samples For TSPI columns
for i = 1 to number of data points of UUT do
  for j = 1 to number of TSPI columns (10) do
    if i modulus idx[j] then
      abs_error[i] ← |uut[i, j] - ref[i, j]|
    end if
  end for
end for
for j = 1 to number of TSPI columns (10) do
  Stats[j, 0] ← RMS of abs_error[j]
  Stats[j, 1] ← Standard deviation of abs_error[j]
  Stats[j, 2] ← number of data points in abs_error[j]
end for
```

---

**Figure 3.** Algorithm 3 (Source: N. Ruprecht).

---

**Algorithm 4** Step 4 Pseudocode - Run Hypothesis Test and Print Results

---

```
specs ← target specification values of unit under test
t ←  $\frac{\text{Stats[:,0]} - \text{specs}}{\text{Stats[:,1]} / \sqrt{\text{Stats[:,2]}}}$ 
p = Python stats.t.sf(|t|, Stats[:, 2] - 1)
for i = 1 to number of TSPI columns (10) do
  if p[i] < α and t[i] < 0 then
    Reject  $H_0$ 
  else
    Fail to Reject  $H_0$ 
  end if
end for
```

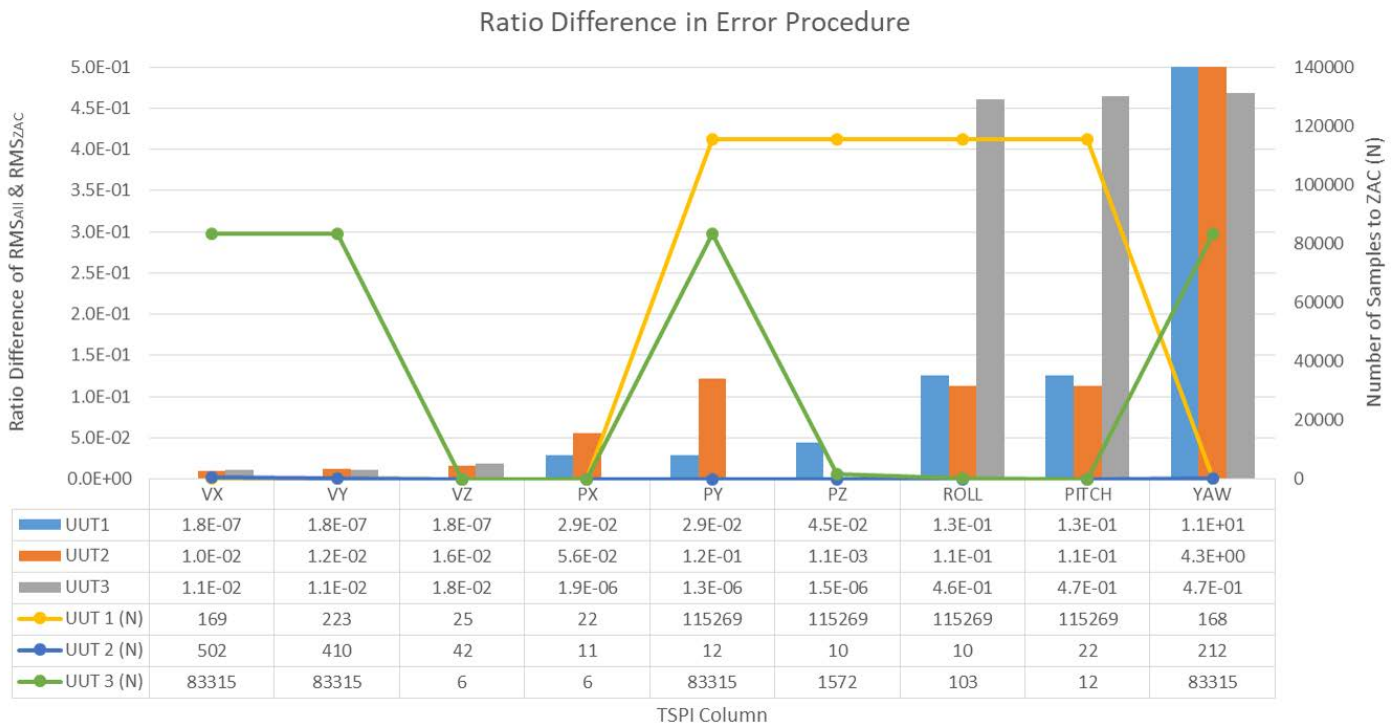
---

**Figure 4.** Algorithm 4 (Source: N. Ruprecht).

to have a starting assumption that the UUT error is too great or considered “out of spec.” When using a hypothesis test, the null hypothesis cannot be “accepted”—only “fail to reject” the null. The UUT would rather be proven to be within specification by rejecting the current null hypothesis, as opposed to fail to reject the idea of it being within specification. Failing to reject the null hypothesis can be interpreted as meaning that the UUT has a greater error than required or that not enough data was collected to statistically prove the error is less.

## RESULTS AND DISCUSSION

Raw errors and results are summarized next. Figure 5 shows the difference in the ratio between the techniques used. The Y-axis on the left corresponds to the difference comparing calculated RMS error when looking at uncorrelated data points vs. error and using all available data. Due to the different scales of magnitude for each TSPI column unit, a ratio difference was used to standardize an output for visual representation’s sake. The Y-axis on the right aligns with the number of samples used for each UUT and TSPI column to achieve a statistically significant (p-value <0.05) ZAC for RMS calculation and evidence-based decision making in the number of data points needed for each column of interest.



**Figure 5.** Calculated RMS Using All Available Data ( $RMS_{All}$ ) vs. Calculated RMS With Uncorrelated Values or ZAC ( $RMS_{ZAC}$ ) (Source: N. Ruprecht).

The figure combines two thought processes—the difference in calculated error along the left y-axis and the number of data points it took to get a statistically significant uncorrelated error along the right y-axis. For nearly all TSPI columns and UUTs, the difference in error calculation technique is near zero and therefore comparable. The few instances where the differences were larger were consistent across UUTs or TSPI columns. Another interesting finding is how small  $N$  can be and still have statistically significant results. Where 30 samples are a good rule of thumb, this shows that conclusions can be drawn with less while, at times, more is necessary. If a representative dataset is available, this methodology can be

used to estimate the target number of samples and required test time more empirically than as a rule of thumb. Overall, negligible difference leans into the purpose of this article that asserts the methodology presented can be used so that the tester can speak definitively about how much data, flight time, and sorties will be required well before the test execution itself.

When evaluating datasets after the fact, the aircraft could enter and exit degraded environments so that the entire flight is broken down to intervals of interest. Here, the results of an entire flight can be the sum of its parts such that using this methodology on each part will have the same conclusion as running on the entire flight. An interesting outcome to see is

how much the interval and number of samples used varies for the same flight. Looking more into the flight profile, it intuitively speaks to these values.

Since correlation is inversely proportional to the variance of data being used, a repetitive flight such as racetracks or orbits will require more data to have temporal separation to decorrelate. In contrast, columns with higher variance saw the correlation coefficient approach zero much earlier. Variance also plays directly into the results of the hypothesis test. With a higher variance in the flight profile, more data points are collected due to a smaller time interval before decorrelation occurs. This almost equates to a lower variance in the flight profile, therefore collecting



“

*Since correlation is inversely proportional to the variance of data being used, a repetitive flight such as racetracks or orbits will require more data to have temporal separation to decorrelate.*

less data due to a larger interval. The t-statistic is therefore mainly determined by the RMS error.

## CONCLUSIONS

This study presents a foundational methodology for determining the necessary amount of data required to achieve statistically significant navigation performance evaluation results. While this approach has not yet been extensively validated against surveyed systems or multiple reference standards, its potential impact on test planning and execution is evident. The key contribution of this work is demonstrating that by leveraging the correlation coefficient, an optimal sample size can be determined for error estimation without relying on traditional heuristics such as the 30-sample rule.

The results presented here indicate that for nearly all TSPI parameters

and UUTs, the calculated root mean squared error (RMS) using all available data ( $RMS_{All}$ ) and the RMS derived from statistically independent data points ( $RMS_{ZAC}$ ) yield nearly identical values. This suggests that decorrelating the dataset before error computation does not introduce significant bias and can be a viable alternative for test planners. Moreover, the findings highlight that the number of required samples can be highly variable, depending on the variance and structure of the test flight profile. Repetitive maneuvers, such as racetracks or orbits, introduce greater temporal correlation and therefore require a larger dataset to achieve statistical independence, whereas more dynamic flight profiles with higher variance can decorrelate more quickly.

A key implication of this methodology is its ability to assist in defining the required test duration and sortie count before execution. By applying this method to preexisting UUT datasets—without necessitating a reference system—engineers can estimate the minimum number of independent samples needed. This enables more precise test planning, ensuring that sufficient data is collected without unnecessary resource expenditure. Given the constraints of flight test programs, including fuel limitations, airspace availability, and cost considerations, this methodology provides an empirical, data-driven approach to optimizing test efficiency.

## RECOMMENDATIONS

Future work should focus on validating the method presented here against independently surveyed reference systems and assessing its applicability across a wider range of navigation technologies and test conditions. Incorporating additional error sources like measurement noise and sensor drift into the model could further refine the accuracy of the predicted sample size. By establishing a standardized process for determining statistically significant test durations, this methodology can possibly improve the rigor and efficiency of navigation system evaluations across military and civilian applications. ■

## ACKNOWLEDGMENTS

The authors would like to acknowledge and thank the 746 TS and 704th Test Group for providing administrative oversight and support to maintain high standards of test rigor and approachability to improve capabilities for the test community. While specifics could not be disclosed for security reasons, the squadron's and group's commitment allowed free flow of thought and communication that was invaluable for improvement.

# REFERENCES

- [1] Blewitt, G. "Basics of the GPS Technique: Observation Equations." *Geodetic Applications of GPS*, pp. 10–54, 1997.
- [2] Falco, G., G. A. Einicke, J. T. Malos, and F. Dovis. "Performance Analysis of Constrained Loosely Coupled GPS/INS Integration Solutions." *Sensors*, vol. 12, no. 11, pp. 15983–16007, 2012.
- [3] Li, Y., J. Wang, C. Rizos, P. Mumford, and W. Ding. "Low-Cost Tightly Coupled GPS/INS Integration Based on a Nonlinear Kalman Filtering Design." Proceedings of ION National Technical Meeting, pp. 18–20, 2006.
- [4] Wendel, J., J. Metzger, R. Moenikes, A. Maier, and G. Trommer. "A Performance Comparison of Tightly Coupled GPS/INS Navigation Systems Based on Extended and Sigma Point Kalman Filters." Proceedings of the 18th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2005), pp. 456–466, 2005.
- [5] Kim, H.-S., S.-C. Bu, G.-I. Jee, and C. G. Park. "An Ultra-Tightly Coupled GPS/INS Integration Using Federated Kalman Filter." Proceedings of the 16th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS/GNSS 2003), pp. 2878–2885, 2003.
- [6] Xu, B., Y. Wang, and X. Yang. "Navigation Satellite Clock Error Prediction Based on Functional Network." *Neural Processing Letters*, vol. 38, no. 2, pp. 305–320, 2013.
- [7] Ober, P. B. "Integrity Prediction and Monitoring of Navigation Systems," 2004.
- [8] Misra, P., and P. Enge. *Global Positioning System: Signals, Measurements and Performance*. Second edition, vol. 206, 2006.
- [9] Diggelen, F. *GNSS Accuracy: Lies, Damn Lies, and Statistics*. Vol. 9, pp. 41–45, January 2007.
- [10] Schulte, R. J., and D. W. Dickinson. "Four Methods of Solving for the Spherical Error Probable Associated With a Three-Dimensional Normal Distribution." Technical report, U.S. Air Force Missile Development Center, Holloman Air Force Base (AFB), NM, 1968.
- [11] Hogg, R. V., E. A. Tanis, and D. L. Zimmerman. *Probability and Statistical Inference*. Upper Saddle River, NJ: Pearson/Prentice Hall, 2010.
- [12] Van Belle, G. *Statistical Rules of Thumb*. John Wiley and Sons, 2011.
- [13] Rosenblatt, M. "A Central Limit Theorem and a Strong Mixing Condition." *Proceedings of the National Academy of Sciences*, vol. 42, no. 1, pp. 43–47, 1956.
- [14] Singh, A. S., and M. B. Masuku. "Sampling Techniques and Determination of Sample Size in Applied Statistics Research: An Overview." *International Journal of Economics, Commerce and Management*, vol. 2, no. 11, pp. 1–22, 2014.
- [15] Carling, K. "Resistant Outlier Rules and the Non-Gaussian Case." *Computational Statistics and Data Analysis*, vol. 33, no. 3, pp. 249–258, 2000.
- [16] Iacobucci, D., S. Roman, S. Moon, and D. Rouzies. "A Tutorial on What to Do With Skewness, Kurtosis, and Outliers: New Insights to Help Scholars Conduct and Defend Their Research." *Psychology and Marketing*, 2025.
- [17] Wan, X., W. Wang, J. Liu, and T. Tong. "Estimating the Sample Mean and Standard Deviation from the Sample Size, Median, Range and/or Interquartile Range." *BMC Medical Research Methodology*, vol. 14, pp. 1–13, 2014.
- [18] Sahasranand, K., F. C. Joseph, H. Tyagi, G. Gurrall, and A. Joglekar. "Anomaly-Aware Adaptive Sampling for Electrical Signal Compression." *IEEE Transactions on Smart Grid*, vol. 13, no. 3, pp. 2185–2196, 2022.
- [19] Nguyen, A.-T., S. Reiter, and P. Rigo. "A Review on Simulation Based Optimization Methods Applied to Building Performance Analysis." *Applied Energy*, vol. 113, pp. 1043–1058, 2014.
- [20] Razavi, S., A. Jakeman, A. Saltelli, C. Prieur, B. Iooss, E. Borgonovo, E. Plischke, S. L. Piano, T. Iwanaga, W. Becker, et al. "The Future of Sensitivity Analysis: An Essential Discipline for Systems Modeling and Policy Support." *Environmental Modelling and Software*, vol. 137, p. 104954, 2021.
- [21] Fielding, S., P. Fayers, and C. R. Ramsay. "Analysing Randomised Controlled Trials With Missing Data: Choice of Approach Affects Conclusions." *Contemporary Clinical Trials*, vol. 33, no. 3, pp. 461–469, 2012.
- [22] Griffith, D. A., and R. E. Plant. "Statistical Analysis in the Presence of Spatial Autocorrelation: Selected Sampling Strategy Effects." *Stats*, vol. 5, no. 4, pp. 1334–1353, 2022.
- [23] Karunarathna, I., P. Gunasena, T. Hapuarachchi, and S. Gunathilake. "The Crucial Role of Data Collection in Research: Techniques, Challenges, and Best Practices." *UVA Clinical Research*, pp. 1–24, 2024.
- [24] Wang, H., and Z. Abraham. "Concept Drift Detection for Streaming Data." The 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–9, IEEE, 2015.
- [25] Osborne, W., and A. Overbay. "The Power of Outliers (and Why Researchers Should Always Check for Them)." *Practical Assessment, Research, and Evaluation*, vol. 9, no. 1, p. 6, 2019.
- [26] Morey, R. D., R. Hoekstra, J. N. Rouder, M. D. Lee, and E.-J. Wagenmakers. "The Fallacy of Placing Confidence in Confidence Intervals." *Psychonomic Bulletin and Review*, vol. 23, pp. 103–123, 2016.
- [27] Schwarz, N., F. Strack, A. Gelman, S. M. van Osselaer, and J. Huber. "Commentaries on Beyond Statistical Significance: Five Principles for the New Era of Data Analysis and Reporting." *Journal of Consumer Psychology*, vol. 34, no. 1, pp. 187–195, 2024.
- [28] Yang, J.-S. "Travel Time Prediction Using the GPS Test Vehicle and Kalman Filtering Techniques." Proceedings of the 2005 American Control Conference, pp. 2128–2133, IEEE, 2005.
- [29] Hong, S., M. H. Lee, S. H. Kwon, and H. H. Chun. "A Car Test for the Estimation of GPS/INS Alignment Errors." *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 3, pp. 208–218, 2004.
- [30] Brown, A., and Y. Lu. "Performance Test Results of an Integrated GPS/MEMS Inertial Navigation Package." *Proceedings of ION GNSS*, vol. 2004, Citeseer, 2004.
- [31] Amt, H., and J. F. Raquet. "Flight Testing of a Pseudolite Navigation System on a UAV." U.S. Air Force Institute of Technology: ION Conference, 2007.
- [32] Stratton, A. "Flight Test Criteria for Qualification of GPS-Based Positioning and Landing Systems." Proceedings of the 22nd International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2009), pp. 1610–1618, 2009.
- [33] Jacobs, T. "Challenges of Moving a Mixed Technology Manual Test Setup Into an Automated Environment." 2024 IEEE AUTOTESTCON, pp. 1–5, IEEE, 2024.
- [34] Defense Logistics Agency. *Gage Inspection*. MIL-STD-120, 1950.
- [35] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC). *General Requirements for the Competence of Testing and Calibration Laboratories*. ISO/IEC 17025:2017, 2017.
- [36] Tiwari, R., M. Arora, and A. Kumar. "An Appraisal of GPS Related Errors." *Geospatial World*, vol. 5, no. 9, 2000.
- [37] Dmitrieva, K., P. Segall, and A. Bradley. "Effects of Linear Trends on Estimation of Noise in GNSS Position Time Series." *Geophysical Journal International*, p. ggw391, 2016.
- [38] Wall, J. H., and D. M. Bevly. "Characterization of Inertial Sensor Measurements for Navigation Performance Analysis." Proceedings of the 19th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2006), pp. 2678–2685, 2006.
- [39] Mari, D. D., and S. Kotz. "Correlation and Dependence." *World Scientific*, 2001.
- [40] Szekely, G. J., M. L. Rizzo, N. K. Bakirov, et al. "Measuring and Testing Dependence by Correlation of Distances." *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

## BIOGRAPHIES

**NATHAN A. RUPRECHT** is a U.S. Air Force (USAF) active-duty engineer in navigation systems and nuclear physics and assistant director of operations at the 21st Surveillance Squadron, Air Force Technical Applications Center, deploying as a scientist. Previously a flight test engineer at the 746 TS, he is in the process of transitioning to the National Air and Space Intelligence Center. Dr. Ruprecht holds a B.S. and M.S. in electrical engineering from the University of North Texas and a Ph.D. in biomedical engineering from the University of North Dakota.

**ELISA N. CARRILLO** is a systems engineer with Raytheon Technologies, with expertise in validation and verification. Her research interests include cross-functional communication as a key element of systems integration, test, and validation. Ms. Carrillo holds a B.S. in mechanical engineering from the University of Texas at El Paso and is studying for her M.S. in systems engineering at Embry-Riddle Aeronautical University.

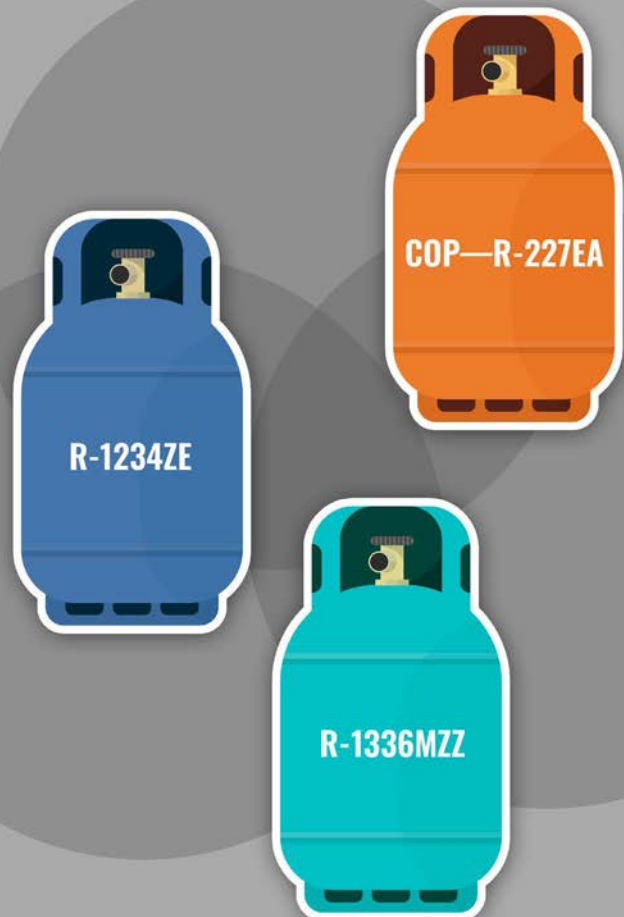
**LOREN E. MYERS** is an active-duty flight test engineer at the Air Dominance Combined Test Force, Edwards AFB, CA, where he executes flight test operations for the F-22A Raptor. His experience includes National Security Space Launch operations and developmental test of navigation systems at the Central Inertial and GPS Test Facility at Holloman AFB. Capt. Myers holds a B.S. in computer engineering from Washington State University, an M.S. in electrical engineering from the Air Force Institute of Technology, and an M.S. in experimental flight test engineering from the USAF Test Pilot School.

## WEB EXCLUSIVE

### INVESTIGATING LOW GLOBAL-WARMING- POTENTIAL REFRIGERANTS FOR MILITARY TRANSPORT APPLICATIONS

By Claire E. O'Malley, Changkuan Liang,  
Enrique A. Velazquez, Robert E. Ferguson,  
Listier A. Otieno, Steven F. Son, and Davide Ziviani  
*Photo Source: Canva*

Results from the study presented in this article will provide the DoD with safe and environmentally conscious refrigerant-blend options.



**AVAILABLE ONLY ONLINE**

<https://buff.ly/bDrIj7h>



# TECHNICAL INQUIRY SERVICES

## FOUR FREE HOURS

Research within our 10 focus areas available to academia, industry, and other government agencies. Log in to <https://dsiac.dtic.mil> to submit your inquiry today.

## TECHNICAL AREAS

Survivability & Vulnerability  
Advanced Materials  
Autonomous Systems  
Non-Lethal Weapons  
Weapons Systems  
Military Sensing  
Directed Energy  
Energetics  
RMQSI  
C4ISR

Photo Source: 123rf.com, SURVICE engineering, U.S. Marine Corps, U.S. Navy, and U.S. Army



The Defense Systems Information Analysis Center (DSIAC) is a component of the U.S. Department of Defense's (DoD's) Information Analysis Center (IAC) enterprise, serving the defense enterprise of DoD and federal government users and their supporting academia and industry partners.

[HTTPS://DSIAC.DTIC.MIL](https://dsiac.dtic.mil)

CONNECT WITH US ON SOCIAL MEDIA

