2022-05-02

Two Gaps That Need to be Filled in Order to Trust AI in Complex Battle Scenarios

Nagy, Bruce

Monterey, California. Naval Postgraduate School

SYM-AM-22-066

Excerpt from the Proceedings of the Nineteenth Annual Acquisition Research Symposium

Acquisition Research: Creating Synergy for Informed Change

May 11–12, 2022

Published: May 2, 2022

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School

The research presented in this report was supported by the Acquisition Research Program at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School

Two Gaps That Need to be Filled in Order to Trust AI in Complex Battle Scenarios

Bruce Nagy—is an applied Research Engineer at the Naval Air Warfare Center, Weapons Division, at China Lake, California. His focus is on integrating game theory and machine learning to create recommendation engines that handle complex, highly dimensional battle management scenarios. In the Navy, Bruce served as an Engineering Duty Officer, where he developed statistically-based algorithms to enhance satellite communications. He was also a technical troubleshooter for the NRO and received honors for his efforts in recovering a failing national defense program that jeopardized U.S. global security. Nagy is also driving a joint effort to establish guidelines for an appropriate level of rigor to be used during the creation of AI-enabled systems. He has received four degrees in math, science, and engineering, and he has conducted postgraduate work analyzing brain stem to muscle group neurology. [bruce.m.nagy.civ@us.navy.mil]

Abstract

In human terms, trust is earned. This paper presents an approach on how an AI-based Course of Action (COA) recommendation algorithm (CRA) can earn human trust. It introduces a nine-stage process (NSP) divided into three phases, where the first two phases close two critical logic gaps necessary to build a trustworthy CRA. The final phase involves deployment of a trusted CRA. Historical examples are presented to provide arguments on why trust needs to be earned, beyond explaining its recommendations, especially when battle complexity and opponent surprise actions are being addressed. The paper describes discussions on the effects that surprise actions had on past battles and how AI might have made a difference, but only if the degree of trust was high. To achieve this goal, the NSP introduces modeling constructs called EVEs. EVEs are key in allowing knowledge from varying sources and forms to be collected, integrated, and refined during all three phases. Using EVEs, the CRA can integrate knowledge from wargamers conducting tabletop discussions as well as operational test engineers working with actual technology during product testing. EVEs allow CRAs to be trained with a combination of theory and practice to provide more practical and accurate recommendations.

Introduction

What does trust in Artificial Intelligence (AI) mean? October 2020, Sandia National Laboratories (SNL) conducted a “Trusted Artificial Intelligence” roundtable with national leaders and industry experts. According to SNL, “AI is trusted if its output can be used in key decision making, including cases where lives may be at stake” (Sandia National Labs, 2021). In a May 26, 2021, memo outlining DoD Plans for Responsible Artificial Intelligence, Deputy Secretary of Defense Dr. Kathleen Hicks stated:

> As the DoD embraces artificial intelligence (AI), it is imperative that we adopt responsible behavior, processes, and outcomes in a manner that reflects the Department’s commitment to its ethical principles, including the protection of privacy and civil liberties. A trusted ecosystem not only enhances our military capabilities, but also builds confidence with end-users, warfighters, and the American public. (Hicks, 2021)

Trusted AI has quite a high bar to pass. Not only must AI be trusted in cases where lives may be at stake and the mission at risk, it must also have the confidence of the American public.

Trust in AI, like trust in human relationships, takes time to build. AI must first prove itself in multiple iterations of complex, realistic synthetic battle scenarios before it can be trusted in actual conflict. The American public needs to know that it’s been thoroughly tested, evaluated, and validated before it’s used to place their sons, daughters, husbands, wives, fathers, and

mothers in harm's way. AI reliable performance can also be considered a system safety issue involving hazards that can embarrass the U.S. (Nagy, 2021).

The degree to which AI should be employed in a complex battle scenario is directly proportional to the degree of trust in that AI system. In other words, the greater the objective, the greater the requirement for trust. For more important decisions, where lives may be at stake, it should require an even higher level of trust, but it also depends greatly on decision-makers weighing risk versus the reward for using AI. For instance, an opponent seeking to get inside an adversary's decision cycle might leverage AI to the degree that he or she determines that the reward outweighs the risk. Likewise, an adversary seeking to disrupt an opponent's decision cycle may want to increase their perceived risk, or choose to mistrust their AI systems.

There are many sayings regarding the need to "earn trust," or the need to "build a relationship based on trust," and an AI system making potential life and death recommendations should not be an exception. This paper will provide a nine-stage process (NSP) describing an approach that creates a trustworthy COA recommendation algorithm (CRA), proven through reliable performance in various functional roles. This approach helps to ensure that decision-makers can be confident in the COA recommendations, especially when life is at stake or an adversary is actively working to create mistrust. Will the forecast of surprise attacks and/or recommendations to commit military resources be trusted? To answer this question, the NSP provides a process that allows the CRA to: (Gap 1) learn tactics and strategies through professional wargaming via tabletop discussions, and (Gap 2) analyze performance limitations and strengths with greater statistical accuracy of products (resulting from technology development/acquisition) via "live" operational testing within Test and Evaluation, Verification, Validation/Live Virtual Construct (TEVV/LVC) facilities. The paper will present why these gaps need to be filled to adequately build trust in a CRA providing critical recommendations within complex battle scenarios.

CRA Tasking in Wargames (Gap 1)

A CRA needs to be developed from a wargaming environment to capitalize on a "treasure trove" of move-to-counter-move knowledge and possibilities, such as: (1) human factors that can affect outcomes, (2) unanticipated/surprise moves changing battle results, (3) multi-domain scenarios, where joint and coalition forces are integrated to achieve a common goal (DSB, 2015), and (4) the ability to accurately interpret various qualities of intelligence/sources. A CRA needs to learn how to unravel battle complexity, including uncovering and managing "unknowns" (DSB, 2009), and still determine an optimal strategy/tactical response. Uncovering "unknowns," meaning revealing surprises in battle before they happen, is challenging. In terms of AI systems, describing "unknowns" within complex battle scenarios, as well as how they can be uncovered or countered before the event, will be reviewed when discussing Event-Verb-Event (EVE) chains and modeling of wargames. A CRA must collect move-to-counter-move knowledge and possibilities, and learn from wargaming experts in order to provide recommendations that can be trusted.

Dr. Peter Perla defines a war game as "a warfare model or simulation whose sequence of events is interactively affected by decisions made by players representing opposing sides, and whose operation does not involve the activities of actual military forces" (Perla, 1987). Perla goes on to state, "The true value of wargaming lies in its unique ability to illuminate the effect of the human factor in warfare. By their very nature, war games seek to explore precisely those messy, 'unquantifiable' questions that campaign analyses ignore. War games can help the participants discover what they don't know they don't know," (Perla, 1987). Wargames, exercises, and campaign or operations analysis are all useful tools to build AI trust. However, exercises tend to be costly endeavors with scripted timelines and campaign analysis is often

bound by analytical frameworks. Only wargames “allow for the continual adjustments of strategies and tactics by both sides in response to developing results and events not seen in campaign analysis” (Perla, 1987). Only iterative wargames over time, with opposing blue and red team members, can render insights into future conflicts.

CRA should be able to use varying levels of intel reliability about the opposing side. Based on intel and performance knowledge of its own technology, it should show strategic and tactical bottlenecks, strengths and weaknesses, as well as ways to improve self-resiliency to ensure success of mission from both red and blue perspectives. Additionally, CRA should be able to adjudicate red and blue moves and countermoves. The CRA design needs to store complete wargames, with details, and then stochastically reenact the wargame to collect statistical results for analysis. It should also be able to alter the levels of intel reliability for either opponent, and through additional analysis show trends and variations. The CRA needs to provide support of blue, red, and white players in three ways:

1. Run the wargame from a blue perspective: It needs to run the wargame from a blue perspective (with related allies) based on what blue “thinks” red (and related allies) will do; but use red “true” actions and intent during the game. In other words, it’s just a blue teammate experiencing, with its blue team members, how well it anticipated red actions. This is analytical assessment from a blue teammate perspective. The CRA needs to learn and share those statistical results with blue team members regarding how to prepare better for unanticipated, “out-of-the-box” surprises in battle from a blue perspective.
2. Run the wargame from a red perspective: CRA performs that same tasking as in the previous step, just from a red perspective. This is considered "playing red with fidelity and rigor" (Rielage, 2017). The CRA needs to learn and share those statistical results with red team members regarding how to prepare better for unanticipated, “out-of-the-box” surprises in battle from a red perspective.
3. Run the wargame from a white perspective: The CRA needs to adjudicate red and blue team moves. It needs to simulate “what if” scenarios. This is an analytical assessment from a white cell player perspective, knowing performance truth of blue and red teams. The CRA needs to provide the blue team with a move-by-move analysis on how intel and/or technical capability were used, skewed, hidden, or even missed, and how those decisions impacted results. The CRA needs to do this for blue and red, using the white perspective to help white team players during debriefs of games. The CRA role needs to be constantly assessing patterns from thousands of seemingly uncorrelated data to learn how to minimize impacts from unanticipated, “out-of-the-box” surprises when making recommendations to its blue and red teams for the next set of wargames. Analyzing thousands, potentially millions of data points, should be a natural CRA function to perform. These types of analyses would be directed by white cell players to discover strategic and tactical bottlenecks, strengths, and weaknesses, as well as to improve resiliency of the systems completing the various missions.

The CRA needs to be able to analyze past wargames to support comparisons of tactical structures based on intel as well as to provide options during evaluation. As more data is collected, the more the CRA is able to provide statistical evidence, segment by segment, following the NSP. It can then use this evidence to make recommendations to blue and red teams on how to be more effective in achieving their mission goals. For the white team, it would automate adjudication, saving wargamers’ time and allowing for more statistically precise analysis of moves. This would start to fill the first gap by providing value to red, blue, and white players in support of their goals and work habits.

A final reason to have the CRA integrated into the wargaming environment is complexity. Does the training data sufficiently represent the deployed challenges involved with battle complexity (Nagy, 2021)? If not, how can a COA recommendation be trusted when battle complexity is experienced during operational deployment? If trust is being earned, then the CRA must make recommendations, with a high statistical likelihood of success that consider battle complexity challenges. During a complex battle engagement, when functions affecting loss of life, property, or key objectives hang in the balance, trusting an unproven/inexperienced CRA in battle, even with human oversight, seems unreasonable. Yet, the motivation for deploying a CRA includes faster reaction time, avoiding human loss, and if trained properly, greater precision in action or detail in recommendation. Therefore, the issue becomes the meticulous process of training or evolving the technology into a trustworthy CRA. The only way this training can be successfully demonstrated is by having a high percentage of successful outcomes when following the recommended COA.

Variables that need to be considered: Does training data adequately prepare a CRA to reliably perform when challenged with the complexity of battle? Will the wargames include the proper complexity? Complexity consists of four elements that when combined make something complex: adaptability, interdependence, interconnectedness, and diversity. According to the scientific definition of complexity, a problem is more complex if it has more of these characteristics (Frank, 2015). Complexity theory is command and control theory: both deal with how a widely distributed collection of numerous agents, acting individually, can nonetheless behave like a single, even purposeful entity (Schmitt, 2008). Most times in literature, the definition of battle complexity can be summarized as a situation where there are many military components, systems, and subsystems interacting for a single purpose against an equally complex opposing force. Note: the term “complicated” relates to difficulty. Dr. Bonnie Johnson and CAPT Scot Miller, U.S. Navy (ret.), both from the Naval Postgraduate School, and other research scientists have written papers and lectured expressing “unknowns” or “uncertainties” being core to complexity (Johnson, 2019; Logan, 2009).

CRA Tasking in Operational Tests (Gap 2)

A logical next step is to have the CRA move from wargaming tabletop discussions to working with actual “live” operational testing of new technology products being developed/acquired by Department of Defense (DoD) programs. It is important that the CRA learn from firsthand experience what products can and cannot do. This data can then be used to refine the moves and countermoves discussed during the wargaming exercises. This also ensures accuracy in the recommendation. When the CRA has demonstrated reliable COAs based on the guidance of wargamers, the CRA then needs to refine its knowledge using “live” data.

In support of TEVV/LVC facilities, the CRA also needs to be designed to write test scripts that can more accurately identify the strengths and weaknesses of new technology products being developed/acquired by the DoD. By analyzing how well the new DoD technology performs when challenged by the test script scenario, the CRA offers additional value to the wargamers while supporting the operational test engineers.

To be valued, it must provide an automation capability to reduce time in developing test scripts and ensure adequate coverage of requirements. It must also share the analytical and statistical knowledge gained through wargaming to support the operational test engineers in developing more tactical and strategic battle complex test scripts. Performance data can then be used for future wargaming, allowing for any tabletop corrections regarding product performance, and thereby adding additional value to professional wargamers. These types of operational test scripts produced by the CRA will satisfy two goals:

(1) To understand how well developmental AI or any new technology can handle the unexpected, i.e., surprises, in terms of performance, capability, and resiliency
(2) Allow CRA to bridge the gap between what is operationally tested and how it is used in a professional wargaming environment; this added value earns "credits" with regard to trust

Again, from a common-sense standpoint, before the CRA is deployed in an operational environment, the goal is to provide the CRA with performance results from following its created test script scenario. The performance results would include working technology supporting the operational test, as well as the product being reviewed for release. Additionally, it's important the CRA can refine, modify, and even correct assumptions/performance data originally described in the wargaming exercises. It must learn as much firsthand knowledge about working/deployed technology being used in the operational theater as possible. It is vital to include this practical performance knowledge in the training of a CRA expected to provide trustworthy recommendations when deployed.

Bridging the gap between wargamers and operational test engineers is important to consider. How well do professional wargamers and operational testers share knowledge? Wargamers speak in terms of strategies, tactics, and outcomes. Testers speak in terms of requirements, performance capabilities, and statistical results. Do operational test script scenarios adequately or correctly reflect how wargamers' scenarios use those technology/product assets when games are played out? The CRA, following the NSP, ensures this alignment. The paper will demonstrate how the NSP, by using EVE modeling, will align these two domains and allow the CRA to produce cohesive and trustworthy recommendations.

The Certainty of Surprise in Battle Engagements

In a 1955 news conference, President Dwight D. Eisenhower stated, "Every war is going to astonish you in the way it occurred, and in the way it is carried out" (Eisenhower Library, 2022).The current conflict in Ukraine doesn't appear to challenge this assertion, even with six plus decades of new technology. In a 2018 Center for Strategic and International Studies (CSIS) report titled ~~*Avoiding*~~ *Coping with Surprise in Great Power Conflicts*, Mark F. Cancian concludes that "surprise is inevitable" (Cancian, 2018) but also points out four different types of surprise: strategic, technological, doctrinal and political/diplomatic. The report analyzes each type of surprise in detail, making clear that not all surprises are a result of adversary action. Doctrinal surprise, according to Cancian, "is the use of known capabilities or technologies in unexpected ways that produces powerful new effects. Doctrinal surprise can also come from the unexpected failure of our own warfighting concepts" (2018). A recent example of doctrinal surprise is the 2020 Armenian–Azerbaijan conflict in which the Azeri used armed UAS to catch the Armenians by surprise and tipped the scale of conflict in favor of Azerbaijan (Canadian Army, 2021). Trusting AI/machine learning (ML) requires these systems account not only for an adversary's surprise but also when our systems, processes, and procedures unexpectedly fail to work as advertised.

From the CSIS article, Cancian defined surprise as "when events occur that so contravene the victim's expectations that opponents gain a major advantage" (Cancian, 2018). This definition of surprise is too broad for an AI system designed to measure individual bits of data. In AI terms, s*urprise* may result in an opponent gaining a major advantage but the origin of surprise may come from a completely unexpected event, or the cumulative effect of many little *surprises* causing deviation beyond toleration in the AI algorithm. This paper does not address whether or not AI may eventually remove the element of surprise from warfare, although, in the authors opinion that is unlikely. Rather, this paper analyzes what is required for AI to be trusted in future conflicts with the element of surprise ever present.

Surprise is inherent in warfare—considered unbound data—this should be considered a given fact. Another given fact is that AI battle decision aids have been known to "catastrophically" fail when presented with unbounded information (Moses, 2007; Cooter, 2000). How is this problem addressed in AI? Ensure the data sets used to train the AI system accurately reflect the deployed operational state! Ergo, the need for extensive wargaming and operational testing before deployment. Filling these two gaps are not optional; they are required to ensure trust in the CRA.

Lessons Learned on why CRAs Need to Earn Trust Before Operational Deployment

History is replete with examples of the United States being surprised (DSB, 2009), including Chinese entry into the Korean War, North Vietnamese offensive during the Tet holiday, Egyptian and Syrian attacks on Israel in 1973, the fall of the Shah in 1979, the fall of the Berlin Wall in 1989, terrorist attacks on September 11, 2001 (Cancian, 2018), and the tenacity of Ukrainian civilians to stand up to a Russian attack on their homeland.

Could a CRA have predicted the December 7, 1941, attack on Pearl Harbor? If so, would leadership have trusted the prediction? Cancian (2018) points out that the attack on Pearl Harbor was predicted—the problem was a lack of trust in those predictions. It seemed implausible that the Japanese would attack knowing the likelihood of bringing the United States into the conflict. More importantly, what would be needed for leadership to change existing battle plans, dedicate resources, and spend the needed operational funds? Given the strategic location of Pearl Harbor and critical vulnerability of point of loading (POL) logistics, it is very plausible that wargamers predicted this possibility. Cancian (2018) also points out, "The United States had broken the Japanese diplomatic code (MAGIC) and therefore had extraordinary insights into Japanese thinking and intentions. Nevertheless, for a variety of reasons—tight controls over access, gaps in information, delays in transmission, confusion about meaning, preconceptions about where an attack might occur—this extraordinary trove of data was not adequate to alert U.S. forces." If wargamers could have predicted this attack, could an AI decision aid tool have made the same prediction? Even if the prediction was not considered credible, would an optimized, trusted resiliency plan have made a difference (DSB, 2009)?

Looking at the Allied Island-hopping campaign in WW2, there are some battles where AI decision aid tools would have been surprised and probably not effective. Using AI, battle loss might have been minimized with an optimized, trusted resiliency plan. In the best case scenario, the AI decision aid tool would have brought years of both wargaming and operational testing experience to the battle commander. Could that have been used to improve Allied warfighting effectiveness in World War II, both from move, countermove recommendations, as well as resiliency plans in case of surprise events?

Explainable AI May Not Be Enough When Significant Change Is Needed

Using "what ifs" to examine the Pearl Harbor attack from the perspective of the United States (who lost the battle), a question to consider is, could a CRA following the NSP have made a difference in the outcome?

- What if the CRA had participated in running wargames and testing of technology involved with surprise attacks at 3rd Fleet/Pearl Harbor? Would it be able to identify variations in defense preparation, resiliency plans, and tactical recommendations? If so, that data could be used to explain a recommendation to commit forces for a surprise attack. Would the recommendation have been followed using this data?
- What if some of the variations included statistical likelihoods, minimal defense postures, and pattern recognition of Japanese force movements and had been forecast through

the CRA's training process? How much of a difference would explaining these details have helped in getting people to prepare?

- What if the CRA, from running the wargame over and over, learned how to minimize response time to deal with a surprise attack, possibly by ensuring resiliency as defensive preparation, or provide a core counterattack that minimized impact? How many casualties could have been avoided? Yet, with this explainable data, would U.S. leadership have listened?
- What if the CRA had earned a trusted relationship with its Pearl Harbor decision-makers, maybe during professional wargaming events or at test facilities? In other words, the CRA had already impressed its users by giving reliable recommendations in wargames and/or provided analysis that created more effective use of technology to achieve mission results. Given this proven track record, would it have made a difference to the users in choosing to follow its recommendations?

An interesting point regarding this "what if" scenario is that there was a plethora of data regarding the Pearl Harbor attack, but the base cadre did not react. An overwhelming amount of data pointed to a surprise attack. Someone reviewed the data and concluded that that there was a likelihood for a surprise attack. But, no one believed the accumulated data enough to support a commitment of military resources. The lesson learned may be that explainable AI, the human reviewing the data acting as an AI equivalent, was likely not enough. Would coming from a computer have made a difference? The definitive answer is performance history!

The goal is to build a trusted relationship with the CRA. Yet, trust is earned through performance reliability, i.e., it generates a high degree of successful recommendations. It's earned by being a reliable wargaming recommendation tool that has demonstrated its ability to counter unknown-unknowns. It can also earn trust through value-added knowledge from operational testing. Following the NSP, the CRA can develop and earn a positive reputation! Consider that without this proven reputation, no matter how explainable, would a recommendation from an AI algorithm be considered reliable enough to commit sizeable amounts of military assets? The point is that a CRA needs to build its reputation based on performance to earn trust!

If this conclusion has credence, the Pearl Harbor lesson is significant. Even if the perfect AI recommendation system is developed, without past history of trust, explainable AI is not enough. Even if the AI explains its recommendation using past history/training data, for example, that the Japanese attacked the Russian's Port Arthur in China about half a decade ago, explainable AI would not be sufficient. (This history was well known at the time.) The point is that as explainable/historical as the recommendation might be, without trust based on a proven track record, the result of the attack on Pearl Harbor would likely have remained the same. Without a trust history, a military commander is not likely to commit a sizeable number of military resources based on a machine's recommendation. Additionally, a resiliency plan recommended by the CRA, even if it was perfect, may have suffered the same fate because it lacked a history of trust.

Consider the Battle of Midway from the perspective of the Japanese who lost the battle. Would an AI recommendation algorithm have made a difference in that outcome?

The Japanese knew that they had superior forces, more experienced pilots, better aircraft, and an element of surprise. In a wargame that calculates the odds of winning, the Japanese likely determined that they would emerge victorious nearly 100% of the time. As a result, the "unknown-unknowns" for the Japanese significantly affected the outcome of the battle, i.e., they were wrong. They did not account for the Americans breaking their code, which is a surprise in technology capability. Their calculations did not account for the heroic and nearly

suicidal efforts of many naval aviators—a human factors surprise. Japan assumed that rearming and refueling, dangerous operational tasks, would always be handled with utmost care, meaning taking time, but they favored speed, another human factor surprise. Again, consider the "what if" list and whether a CRA could accumulate enough knowledge to uncover these unknowns. If so, would the recommendations have been trusted? If these unknowns could have been uncovered, would the CRA have been believed regarding recommendations to counter these surprise events? Again, this is a relationship issue, earned through performance reliability, i.e., to earn trust, a high percentage of successful recommendations need to be made in wargames. Yet, to make accurate recommendations, the CRA also needs to understand the performance strengths and limitations of products/technology from firsthand knowledge, learned during operational tests.

As described in the Pearl Harbor Attack and the Battle of Midway, interpretation of the intelligence and human factors plays a major role in action and reaction, move and countermove, eventually leading to a final outcome. Maybe the surprises could not have been predicted, but what if the CRA provided a resilience plan that effectively countered the impact of the surprise? As a lesson learned, intel and human factors need to be included in the training of the CRA and the evaluation of its reliable recommendations, both a proactive counter and/or an effective resilience strategy.

Trust to Overcome Hubris May Be the Best Approach

If not considered, human factors can be a surprise element during a wargame. Hubris can adversely affect a rational decision, and trust might be the only human factor that can create needed clarity. How much trust is enough to overcome hubris?

What are the lessons learned when hubris plays a role in making decisions? This question is important to consider because the United States is considered a "superpower." Can the hubris make a CRA recommendation even harder to accept? Does it raise the bar regarding how much trust needs to be earned to overcome hubris for the recommendation to be accepted? Can hubris become a weakness, impacting battle outcomes? Was hubris a major factor in the lack of reaction to overwhelming data stating the Japanese was about to attack Pearl Harbor? Did the Japanese demonstrate hubris during the Battle of Midway attack?

A potential example of hubris might be in an interpretation of the Battle of Nagorno–Karabakh War from the perspective of the Armenians who lost the battle. This is purely conjecture regarding attitude and must be emphasized that this discussion is being provided as an example only. This interpretation may be false, but will be used to emphasize a point regarding the potential that hubris may make it more difficult to trust an AI system providing CRA recommendations. The conjecture is that the Armenians assumed a weaker opponent and although intel stated a buildup of capability across the border, hubris of their past success overruled their caution. The result is that Azerbaijan actually proved themselves during battle to be a peer adversary. This was a surprise to the Armenians.

Armenian confidence was based on a history of success with Azerbaijan, considered "known-knowns" (past history). Azerbaijan confidence was based on increased "known-unknowns" (assumptions about improvements). The Armenians won the last war and thought they would win the next (assumptions). They were not prepared for Azerbaijan's improved battle capability (surprises). On the other hand, Azerbaijan learned from the last war. They increased their technology and military training by linking with winning Russian technology and strategies. As a result, they were able to significantly alter the Armenian's expected outcome of the battle. Azerbaijan's had Russian's Snowdome defense and Armenian's did not anticipate its effectiveness. Was this poor intel or hubris? This was a technology surprise factor. The effectiveness of UAS (used in good weather) and tank artillery (used in bad weather) severely

reduced Armenian capability—additionally, use of the Israeli Harop (UAS) to provide both surveillance and kinetics was effectively used, another Azerbaijan technology surprise.

Would a CRA, trusted through proven recommendation performance through wargaming and operational testing, have enough earned reputation to overcome any potential Armenian hubris? Like in the U.S. example where intel pointed to a Pearly Harbor attack, intel data was not sufficient. With regard to the United States, does its status as a superpower cause hubris among its military leadership, and would a CRA with a proven track record in wargaming and operational testing (e.g., understanding the performance effects of Snowdome or Harop) have made a difference? Would AI explainability of the data used to make the recommendation be enough? How much trust would have been needed, along with explainable data, to convince Armenians that they needed to better prepare?

Avoid Designing a CRA to Earn Limited Trust

Another human factor to consider is a willingness to die for one's belief. This has been true with suicide bombing. Suicide bombing is a surprise tactic that has occurred in Arabic wars, as well as during World War II. Japanese Kamikaze bombing was completely unanticipated. From the perspective of the United States, could this have been foreseen? Could this sacrifice have been anticipated by a CRA? The first kamikaze pilot to drive his airplane into a WWII warship likely would have been a significant departure from predicted norms, and a surprise to any modern-day AI system. Could the CRAs have dealt with kamikaze attacks? It is possible a modern-day AI system could have analyzed Bushido code (Anya, 2013) to recognize that Japanese culture placed a great deal in sacrificing life for honor and from that made a correlation to the possibility of future kamikaze attacks. The solution is obviously valuable, but does it earn trust?

This correlation relies on AI programmers to input the Bushido code to support a kamikaze prediction. This type of training data is considered bias. The challenge is that the variance between data sets would be very poor, meaning that the data would not support other cultural relationships from other countries, e.g., suicide bombers following a different religious code. As a lesson learned, it is important that CRA training avoid this bias limitation. This paper is not recommending excluding this approach, but ensuring the trust is not dependent on it when dealing with opponent surprises. If it is, then the CRA would be limited to Japanese actions associated with the Bushido code. The trust would be earned within this realm, but not others. As will be described, CRA needs to be a generalized, structured approach to deal with the wide variety of opponent surprises. This is important if the CRA is to be trusted, i.e., bias and variance should be balanced.

As an alternative, generalized approach, a kamikaze attack may not have been anticipated, but the CRA may have considered how to deal with specific types of impacts from opponent moves and countermoves. This is the benefit from using EVEs. The CRA may also have developed a resiliency plan based on assuming opponent success, another benefit from EVE modeling, thereby minimizing the effects of the opponent impact. Recommended readiness and resiliency, especially coming from a trustworthy CRA, is a proven defense against the unknown. The key is having the CRA understand vulnerability points and to make trustworthy recommendations for countermoves that include resiliency. Training of the CRA to understand vulnerability and recommend the needed response is provided using EVE Chains.

The Power of EVE Chains for CRA Development

Data needs to be collected during wargaming and operational tests as part of the training process. EVE chains are designed to replicate any type of action or exchange of actions (Nagy, 2021; Nagy, 2022). In wargaming, EVEs can model moves and countermoves based on

the world state. An EVE segment can represent a specific move, countermove interaction, capturing each wargaming interaction into EVE segments for reuse. In operational test, it can represent a sequence of actions required for the product under test to perform, including evasion and other forms of counteractions. The EVE model consists of events, state variable changes resulting from verb execution. State variables comprise an event. The verb modifies certain state variable, thereby creating a resulting event. Here are some common terms used with regard to the EVE modeling:

- Event: All or part of a world state at a specific timeframe – the world state consists of all enabler and influencer state variables involved with game play. In this TAWC construct, events consume no time during game play.
- Verb: An action available to the Enabler and Influencer that changes the world state and consumes time on the game board. Verbs can be functionally represented by certified meta-models (Nagy, 2022), ML algorithms, or polynomials. Note: the combination of EVEs with meta-models allows for lightweight, low-processing power systems proven by the Battle Readiness, Engagement Management (BREM) prototype project (Nagy, 2022).
- Enabler: An asset, a "piece" within the game, that has specific Verbs (or actions) that when performed can affect the world state, e.g., Enabler Verbs can counter the negative effects of entity influencer actions and counter obstructions; or enabler verbs can take advantage of an obstruction that supports mission success. Note: Depending on perspective, Enablers can be blue or red game pieces. Enablers are only represented by Entities.
- Influencer: There are environmental and entity influencers. Environmental influencers consist of moveable and immovable obstructions, as well as weather conditions above and below. Entity influencers have Verbs that when performed can negatively affect one or more Enabler Entity state variables, as well as moveable obstacles in a way that causes mission failure. Note: Depending on perspective, Influencer Entities can be blue or red game pieces.

Training data, for both wargames and operational tests, are EVE chains that can be collected and statistically analyzed. CRA put together EVE chains as recommendations, i.e., a high statistical percentage of successful outcomes, involving actions (verbs) based on input state variables (events that lead to mission success. EVE segments are created from data collected from wargames, i.e., moves and countermoves, as well as product test results, i.e., test scripts demonstrating performance of required moves and countermoves. From wargaming and test operations, EVE segments, moves and countermoves, can be defined and refined from the same pool, supporting greater model accuracy. Greater model accuracy means opportunities for greater recommendation accuracy.

To reiterate, a necessary ingredient in the NSP, or any COA development approach, is to earn enough trust, meaning an extremely high percentage of successful EVE chains that included responses to surprise opponent actions, that if the CRA predicted a surprise event, the user would follow its recommendation resulting in a commitment of military forces. Yet, forecasting an EVE chain that is a surprise, meaning never identified within a wargame or operational scenario, is a difficult challenge.

It should be noted that professional wargames are not about learning to predict the future, nor validate friendly or enemy courses of action, i.e., EVE chains. As Perla (1987) stated, it's about "illumination" and "exploration." For the CRA learning process, wargames provide "illumination" and "exploration" of causality. It provides the medium for causal analytics that support the development of EVE sequences. Those EVE sequences lead to outcomes, based on wargame results. When those sequences are statistically analyzed, then the outcomes can

be associated with a likelihood of success. If basic causal analytics can be learned from wargames, then the CRA, playing as all three team colors, can develop more statistics and EVE segments than by playing against itself. It's this accumulation of EVE segments that will support COA's being prepared to deal with surprises (Nagy, 2022).

The CRA is designed to statistically forecast outcomes based on pattern matching EVE segments accumulated during human play or self-play. It tracks actions to EVE sequences and segments as a pattern matching approach. By pattern matching EVE segments, a statistical forecast of an outcome can be produced. As an example, a person (1) wakes up every day at the same time, (2) makes coffee, (3) gets dressed in professional clothes, (4) gets in a car, and (5) goes to work. This is noticed 75 times, and 25 times the person went for gas. If the first four were provided, statistics would be obvious with regard to going to work or getting gas. But would this result be believed?

What if one of four actions was missing from the input. How would the CRA respond with a recommendation for an outcome? A given state is used to determine an action, but when the state is inaccurate, the challenge is for the CRA to still perform reliably. This is a common "garbage in, garbage out" problem and is considered a Bayesian approach to prediction (Adamski, 2019). The NSP requires the CRA to be designed to use ML generalization to deal with this issue. A surprise is when the person gets a ride from a friend. Maybe the car is in the shop. How does the CRA handle this surprise?

From an EVE chain focus, unknown-unknowns events have two parts:

(1) Part 1 is identifying "what" (one of more state variables in the Event) will be impacted that will prevent Verb (or action) from executing, e.g., blow up fuel depot to prevent planes from refueling, or destroy runway to prevent planes from taking off. Each variable can represent a one for available or zero for not available. This is called a binary EVE chain analysis. The opponent wishes to create zeros, thereby preventing any actions. A single zero in the binary EVE chain analysis can impact the success of a mission.

(2) Part 2 is anticipating "how" the opponent will cause the event (one or more state variables) to be impacted, set to zero, e.g., the process that blew up the depot or destroyed the runway. Was it a suicide bomber, a well-placed bomb, or something completely unanticipated?

From an EVE chain focus, there are two responses to counter the attack:

(1) Part 1 is determining the opponent's "how" and then an appropriate counter (with an EVE chain sequence) to ensure the state variables remain one, thereby ensuring that the EVE chain/sequence can continue to achieve mission success.

(2) Part 2 involves creating a contingency EVE chain/sequence that considers that a counter is ineffective, meaning the state variable to set to zero. And EVE chain must show resilience to the impact, i.e., an alternative Verb that maintains a successful mission outcome.

In an EVE chain, there may be thousands of state variables. The CRA, through the process of learning from wargaming and operational testing, has the time to "crunch" through thousands of state variables that might be targeted. It can assess which state variables have the highest impact, optimal counter response and resiliency plan. It can also assess which variables are least attacked but have highest impact, thereby analyzing and sharing unlikely but impactful vulnerability points. Notice that the CRA may not be able to predict how variables will be attacked, but can anticipate likelihood based on impact and counter strategies, including resiliency plans.

For EVE chains associated with AI systems, battle complexity is defined as a situation that can be described by a series of events, i.e., EVE chains, caused by actions between

opposing participants, where the outcomes can be significantly affected by factors categorized as: (1) “known-knowns” (facts), (2) “known-unknowns” (assumptions) (3) “unknown-knowns” (absent data) and (4) “unknown-unknowns” (surprises; Nagy, 2021).

- “Known-knowns” (facts)—factors that participants depend on as “fact” to win the engagement; these can include own participant’s ISR and C2 technical capabilities, geo-spatial, temporal situational awareness, interoperability, EW effects, human skills, tactical actions and strategy pros/cons. These are EVE chains from data collected from wargames and operational tests.
- “Known-unknowns” (assumptions)—factors that each participant needs to “assume” about variations (of the facts) regarding battle conditions, these can include the third-party involvement, weather forecast, IO, ISR and C2 effectiveness, kinetic and non-kinetic effectiveness, opponent’s attack surfaces and related vulnerabilities, heroism and initiative on all sides, opponent’s priorities, and difficulty in overcoming manmade and natural obstructions. These are assumed variations in EVE chains from data collected from wargames and operational tests.
- “Unknown-knowns” (absent-data)—factors that cause a participant to be “absent of data,” sometimes decision critical info; these factors can include human mistakes, sensor failures, and communication issues. These are missing state variables in EVE chains.
- “Unknown-unknowns” (surprises)—factors that will “surprise” participants during the engagement; these include unforeseen technology and anything not anticipated in the previous three categories. These are EVE chains that have not been identified in any wargame or operational test. The NSP will describe how these EVE chains are addressed using generalization (Stage 9, Phase III of the NSP approach).

In a complex battle, where surprises are certain, how do you provide an algorithm with training data, i.e., EVE sequences, to handle surprises when those surprise are unknown? Consider how the CRA is being developed to support this need through wargames and product testing that include unknowns, i.e., degrees of unbound data. This is the reason why these two gaps need to be filled. This is also the reason why the CRA must play against itself, i.e., self-play, to accumulate EVE segments from a variety of moves and countermoves. The CRA can also determine a resiliency plan, another set of EVE segments, when a state variable is least likely to be attacked but has the greatest impact toward mission success remains flipped. Can enough self-play reduce the number of unbound data issues, meaning surprise events? This needs to be determined, but collecting data from wargaming does help.

Even if “surprise” is a given, it is believed that the number of surprises can be reduced through the wargaming effort, thereby reducing the opportunity for unbound issues. This is another important reason why the CRA must learn from wargamers by capturing those EVE segments. In a 1960 speech to the U.S. Naval War College (USNWC), Admiral Nimitz remarked, “The war with Japan had been re-enacted in the game rooms here (USNWC) by so many people and in so many different ways that nothing that happened during the war was a surprise—absolutely nothing except the kamikaze tactics toward the end [of] the war; we had not visualized those”(Nimitz, 1965). For more than a decade during the interwar period, wargamers at the Naval War College had war-gamed every aspect of a potential conflict with Japan and identified nearly every contingency, and yet the war with Japan still brought surprises.

The “earning trust” challenge will always involve the ability to prepare the CRA to handle unbound data issues, i.e., surprises the opponent might unveil during an engagement. Collecting data from wargaming plays a needed role to minimize those surprises, thereby reducing unbound issues. If surprises, like a kamikaze attack, do occur, the CRA needs to be

prepared to provide resilient solutions along with counter solutions, both represented by EVE chains.

There should always be a concern that a surprise might cause the data input to go beyond that algorithm's variation limits. To address this concern, there must be an approach to ensure oversight against these unbounded solutions (Miller, 2021) and ensure that the CRA is ready to provide resiliency recommendations as an alternative. Guardrails and gates are proven approaches for AI algorithms. Unbound data by its inherent definition means that confidence in the performance/behavior of the AI model cannot be predicted and therefore cannot be trusted. In order to represent a realistic operational set of training data, complexity of the deployed environment needs to be considered; resiliency planning must be immediately available. Again, the CRA attempts to address these considerations through its wargaming and product testing. Given surprise is a given, therefore unbound data is a given, wargaming and operational testing that includes resiliency is a needed ingredient for the CRA to earn trust. From these two environments, EVE segments can be collected and the CRA can be trained to recompose to meet variations in mission challenges. The process of training is the NSP.

NSP in Developing a Trustworthy CRA

The NSP is based on using EVEs to connect all parts of the learning process described in each stage using a common model. Details in Figure 1 show the accumulation of EVEs in each of the three phases. This is how to ensure CRAs make trustworthy recommendations, enabling decision-makers to be confident when life is on the line and a commitment of large military resources is needed. Using EVES, the three phases are connected through nine stages. In all three phases, shown in Figure 1, the CRA results in creating tactical or operational plans. Following the nine stages over three phases of the NSP, a trustworthy CRA can be created. Data collection using EVE modeling bridges knowledge between these wargaming and test domains, creating and refining improved recommendations as preparation for the CRA to be deployed. As a needed result of NSP using EVEs, wargamer and operational tester benefit from increased automation and statistical analysis. This motivates the users to continue to use the CRA in their domains, establishing a value-added approach for all involved.

It is also important to note that the NSP involves the training process to earn trust. Stage 1 of Phase I can occur in parallel with the CRA core development. When Stage 1 and CRA core development is complete, then Stage 2 and on can occur. The CRA must have its core development complete using EVE chains or similar modeling structures associated with world state variables. NSP is based on using EVE chains. An example of a CRA core design using EVE chains is provided in a paper by Bruce Nagy presented at the SPIE conference on Defense and Commercial Sensing (Nagy, 2022).

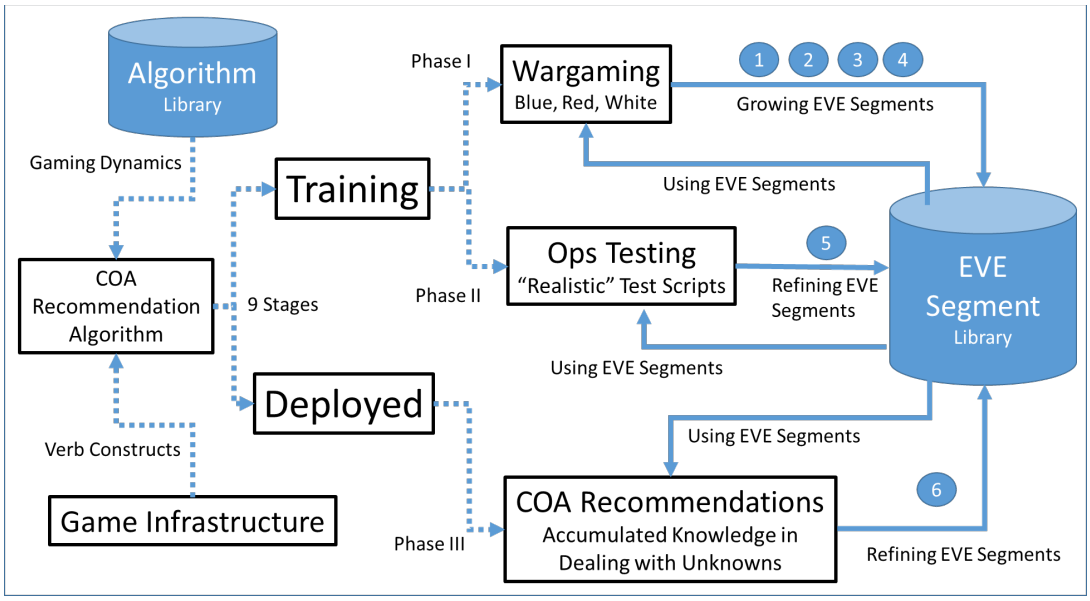

Figure 1. NSP Overview

Phase I Wargaming

The *Wargaming Mode* focuses on supporting professional wargamers. The process involves data collection from professional wargamers using validated performance capabilities of assets and technologies used in games. In this phase, the CRA performs as a wargaming analysis tool acting as a wargaming team member for red, blue, and white teams to support professional wargaming institutes in better analyzing and understanding the effects of intelligence quality, strategy, and tactical outcomes within various wargaming scenarios. There are seven stages within this phase. The first three stages of Phase I are shown in Table 1.

Stage 1 focuses on developing algorithms that move and track game pieces, i.e., assets, on the world board game. It also includes automating various adjudication processes for the wargame users. This is a prerequisite stage and must be done in advance with the focus on developing ancillary algorithm used by the CRA during game play, while also supporting automation needs of wargamers. Although Stage 1 is listed in the wargaming phase, it must also include statistical automation tools that will be used to support the TEVV/LVC facilities. Additionally, this stage establishes all the background information needed to inform the verbs and events in the EVE structure. For instance, if the verb is *move*, and it involves an aircraft entity, stage 1 captures all the performance parameters. In other words, game pieces and moves are automated for war game activity.

Stage 2 determines how the game board is initially set up and what its end goals are. It collects user data that determines what they would consider the beginning states and end states (derived from the commander's intent) for various missions. It sets the stage for the wargame, including placement of assets around the world, their state of readiness, and what goals need to be accomplished. Stage 2 captures the variety of missions, both for blue and red teams. This "current" to "end goal" states can also be entered in "real time," before the game begins. Data entry can be manual or automated about world state for the initial/first event and related state variables, as well as the last/final event, i.e., what the world state needs to "look like" when a mission is concluded. This final/last event supports the commander's guidance translated into world state variables. Notice that when state variables change based on actions, this represents EVE chains (Nagy, 2022). It is not possible to develop credible CRAs if the beginning and end

states are not adequately defined. In this stage, performance bounds are also defined, providing a landscape involved with the game board.

Stages 3 and 4 involve running the CRA using these two previous stages, Stage 1 for automated game piece movement and Stage 2 for game piece placement. This is necessary if the CRA is to optimally determine the best moves and countermoves from each team perspective. Remember that the CRA takes on all team colors involved with game play.

Table 1. Developing CRA Segments 1–3

Phase I Wargaming - Segment 1:	**Phase I Wargaming - Segment 2** (for Blue as Enabler):	**Phase I Wargaming - Segment 3** (for Blue as Enabler):

• Create Verb Infrastructure
  • EVE Ontology
  • Verb Table
  • Verb Binary Codes
  • Verb Hex Code
• Create Algorithm Library
  • Geometry for Global Movement Dynamics
  • Optimization using Learning Rate of Attributes
  • Statistical Significance Analysis
• Create EVE Segment Library (with Sections)
  • Manageable Obstructions
  • Unmanageable Obstructions
  • Enabler Mission Actions
  • Influencer Actions
  • Influencer Counter Actions
★ Establishing EVE Segment Library
	• Define Mission Constraints
	  • Blue Mission Criteria
	  • Blue Performance Area
	  • Blue Environmental Influencers
	  • Immovable Obstacles
	  • Moveable Obstacles
	  • Weather and Other Conditions
	  • Blue Entities
	  • Blue Actual Performance Specs
	  • Blue Allie Estimated Performance Specs
	  • Red Influencer Entities
	  • Red Estimated Performance Specs
	  • Red Allie Estimated Performance Specs
		• Develop Ideal EVE Chain
		  • Movement Dynamics
		  • Performance Area
		  • Immovable Obstacle
		  • Environmental Conditions
		  • Mission Achievement
		  • Tree Trunk and Branches
		  • Verb Stack
		  • EVE Stack (Values)
		  • Binary and Hex Code EVE Stack
		  • EVE Segment Library by Move, Move (Results)
		  • Store Optimal Strategy
		  • Store Statistical Measured Results
		1 Growing EVE Segment Library

Stage 3 is having the CRA develop an optimal strategy and tactics for achieving the end goal state defined in Stage 2 for red and blue teams, as well as their allies. This stage views an ideal world, where opponents and environment influencers are not a factor. It states that if the board did not have opposing pieces or obstacles that it could overcome, what would be the optimal moves to achieve results, i.e., end states. This might generate many solutions that can be analyzed based on team priorities defined in terms of what is considered mission success. From this analysis, the CRA selects "best" candidate(s) with their movement domain route(s) to achieve mission success when there are no opposing/opponent entities. Obstacles may be involved, but limited to those obstructions that cannot be modified, i.e., an immovable landscape.

Stage 4 is having the CRA develop an optimal strategy and tactics for achieving end goal state when opponents forces within a movable and movable landscape, game board. It focuses on having the CRA playout various scenarios between blue and red forces. It attempts to select the optimal game piece candidate(s) with their movement route(s) to achieve mission success where there are opposing/opponent entities attempting to thwart actions. The CRA now has the ability to "go through" environmental/obstructions, if this benefits the users' end state goals and priorities.

Stage 5 is a repeat of stages 2, 3, and 4 but for the opponent. Remember that each team only has limited knowledge, based on the quality of intel about the other player. In other

words, the stages are repeated for both the blue team (with allies) and the red team (with allies) to determine their optimal strategies against each other, not knowing "truth" of the other players' capabilities. Details associated with stages 4 and 5 are shown in Table 2. By completing Stage 4 and 5, best candidates or combinations of game pieces for each opposing team are identified to play out in a non-ideal environment, i.e., opposing/opponent entities and within moveable/manageable and unmovable/unmanageable environmental conditions and obstructions.

Table 2. Developing CRA Segments 4 and 5

Phase I Wargaming - Segment 4	**Phase I Wargaming - Segment 5**

(for Blue as Enabler):
• Develop Non-Ideal (Challenged) EVE Chain
  • Movement Dynamics
  • Performance Area
  • Environmental Influencers
  • Immovable Obstacles
  • Moveable Obstacles
  • Weather and Other Conditions
  • Entity Influencers
  • Blue and Allie Actual Performance Specs
  • Red and Allie Estimated Performance Specs
  • Mission Achievement
  • Tree Trunk and Branches
  • Verb Stack
  • EVE Stack (Values)
  • Binary and Hex Code EVE Stack
  • Counter Moves
  • Verb Stack
  • EVE Stack (Values)
  • Binary and Hex Code EVE Stack
  • EVE Segment Library by Move and Counter Move
  • Store Optimal Strategy
  • Store Statistical Measured Results
2 Growing EVE Segment Library
	(for Red as Enabler):
	• Repeat Segment 2 to 4 for Red Team
	  • Segment 2' (for Red as Enabler):
	  • Define Mission Constraints
	  • Red Mission Criteria
	  • Red Performance Area
	  • Red Environmental Influencers
	  • Red Entities
	  • Blue Influencer Entities
	  • Segment 3' (for Red as Enabler):
	  • Develop Ideal EVE Chain
	  • Movement Dynamics
	  • Mission Achievement
	  • EVE Segment Library by Move and Counter Move
	  • Segment 4' (for Red as Enabler):
	  • Develop Non-Ideal (Challenged) EVE Chain
	  • Movement Dynamics
	  • Mission Achievement
	  • EVE Segment Library by Move and Counter Move
	3 Growing EVE Segment Library

Stage 6 involves having the CRA perform the adjudication process involved with a wargame. This means that the "white cell" runs the wargame with complete knowledge of both sides, red and blue. Their tactics and strategies were based on perception and interpretation from intel sources within the wargame construct. The CRA uses "truth" about capabilities and intent on each side to assess the actual outcome for each side in achieving mission success, given the reality of each side's tactics and strategies. It can then run "what if" scenarios that include variations in performance and intel quality, EVE segment by segment to find optimal outcomes for each side. In game theory, this is finding either the Pure Strategy Nash Equilibrium (PSNE) or the Mixed Strategy Nash Equilibrium (MSNE). These "what if" solutions contribute to various points on a Pareto Analysis chart, i.e., a four-square readiness matrix described in segment 9.

Stage 7 compares the original results from Stages 4 and 5 to Stage 6 modifications, truth at stage 4 and truth at stage 5, comparing perception of the opponent based on intel to the actual truth of the opponent's capabilities. Included in this comparison are the "what if" results. The process is running through each chain sequence, EVE segment by segment, and tracking how often there was an attempt to flip each state variable to zero within the binary EVE segments. The highest number becomes the most likely candidate of vulnerability and the lowest number, the least, given wargaming trends. This stage can be executed/run in the

background or in advance of any wargames, as long as Stage 1 and 2 have previously been completed.

Stage 6 and 7 are shown in Table 3. These stages become a significant learning process for all involved, users and CRA, in identifying how to optimally deal with unknowns, specifically bits that were not flipped and why they were not flipped. Is there a way to create a strategy or tactic that would ensure that a mistake in assumptions has minimal effect on mission outcome? This is what the CRA is being designed to investigate and is unique from other algorithms. From a wargaming, adjudication perspective, the solution can be used for wargaming analysis and adjudication, identifying how and when to adjudicate, and providing unknown-unknown challenges.

Table 3. Developing CRA Segments 6 and 7

Phase I Wargaming - Segment 6:	**Phase I Wargaming - Segment 7:**

• Adjudicate War Game as White Cell
  • Run Monte Carlo Wargame with Existing Assumptions of Blue and Red EVE chains against each other
  • Blue EVE Chain using Assumptions based on Red Intel about Capability
  • Red EVE Chain using Assumptions based on Blue Intel about Capability
  • Statistics by Segment
  • Store for Enabler (Blue) with Influencer (Red) Assumptions
  • Store for Enabler (Red) with Influencer (Blue) Assumptions
  • Run Monte Carlo Wargame with "Truth" of Blue and Red EVE chains against each other
  • Blue EVE Chain using Red "Truth" about Capability
  • Red EVE Chain using Blue "Truth" about Capability
  • Store Statistics by Segment
	• Review Lessons Learned
	  • Comparison of Delta's Assumption vs Truth
	  • Connect Delta's to Statistical Results (Answers Why)
	  • Develop Optimal Solutions based on "Truth" from both sides (EVE Segments with Statistical Results)
	  • Store in EVE segment database by Move, Counter Move (Results from Blue and Red)
	  • Optimal Strategy
	  • Non-Optimal Strategy
	  • Statistical Measured Improvement
	
	4 Growing EVE Segment Library

Phase II T&E

After Stage 2 is complete, Phase II begins the CRA evolvement of earning trust. The T&E Phase focuses on supporting TEVV/LVC facilities. The process involves data refinement from testing technology products and systems using validated performance capabilities of assets and technologies used during testing. In this phase, the CRA performs as a testing analysis tool, a modification of its wargaming capability developed in Phase I. In Stage 8, per Table 4, the CRA is engineered to create rigorous test scripts, while refining EVE segments to better represent "realistic" performance capabilities, results, and limitations. There is only one stage within this phase.

The CRA is now ready for the final upgrade in becoming a recommendation algorithm to generate nominal and stress level test scripts. This is a refinement and validation process from the wargaming EVE segments. Using TEVV/LVC facilities, the EVE segments represent complex environments, linked to live systems or six degree of freedom systems. The CRA

algorithm will adjust based on the performance of all products represented within the environment.

Table 4. Developing CRA Segments 8 and 9

Phase II T&E - Segment 8:	Phase III Deployed Operations - Segment 9:

• Test Script Generator
  • Import Mission Requirements for Test and other Segment 2 Data
  • Mix and Match EVE Segments to create
  • Optimal Solution: (1) EVE Tree with Causal Why, and (2) Why Statistically based on "Truth" of Influencer
  • Nominal Solution: (1) EVE Tree with Causal Why, and (2) Why Statistically based on "Truth" of Influencer
  • Stressed Solution: (1) EVE Tree with Causal Why, and (2) Why Statistically based on "Truth" of Influencer
  • Based on testing, modify EVE segments used to support measured results, including variations in statistics
  • Store Data Changes from Test Results in EVE segment database by Move, Counter Move (Results from Blue and Red)
5 Refining EVE Segment Library
	• COA Recommendation Engine
	  • Import Mission Parameters and other Segment 2 Data
	  • Allow User Preference
	  • Mix and Match EVE Segments to create a Pareto Chart associated with BRE Matrix
	  • Solution given Assumed Intel Truth and Use Preference
	  • Solution given Variations in Intel Assumptions based on Wargaming
	  • Provide Solution that encompasses as many points on the "Green" segment of the Matrix
	  • Not optimal for a single point
	  • Best compromised solution for encompassed points
	  • Store Data Changes from "Live" Operational Results in EVE segment database by Move, Counter Move (Results from Blue and Red)
	6 Refining EVE Segment Library

While developing test scripts and accumulating knowledge, the CRA must be engineered to collect state variables least attacked but with highest impact and not identified in wargames through state variable by state variable investigation. If found, this is considered a paradigm shift, i.e., unknown-unknowns, to support wargamers.

In support of its operational testers, the CRA needs to provide three types of test scripts. Each test script can have subscripts identifying where to change the testing conditions and scenario to support the three types of tests. The three types of test are: (1) nominal performance, (2) product performance under attack and a demonstration of an effective counter, and (3) product performance under attack, not effectively countered, therefore requiring resiliency for the product under review to examine the products limitations. All data is collected and shared with wargamers.

At this point, Phase I and Phase II are being executed simultaneously. Only after both wargamers and test engineers agree will the CRA be allowed to move into Phase III.

Phase III Operational Mode

The operational mode focuses on deploying the CRA to support assets in the field needing to make tactical and battle management decisions. The CRA learning process continues to involve data refinement from operational exercises using live data from assets and technologies. The CRA is an evolvement of the two previous phases. The CRA is now designed

to provide trustworthy recommendations that will also ensure opponent generated surprise issues have minimal effect on the outcome of a mission. There is one stage within this phase.

Stage 9, as described in Table 4, represents the final CRA development stage and a graduation to live operational support**,** i.e., CRA being deployed. From the previous phases, EVE segments have been developed and refined, now available by the CRA when needed. In Phase I, the understanding of intel quality associated with EVE segment selection was analyzed. Additionally, the EVE segments supporting complex environments were created. This resulted in a validation of war gaming complexity and EVE solutions that are statistically significant, meaning mission impacts are unique for each solution. It should also be noted that since EVE segments were developed from professional wargamers using validated technology performance data garnered from real games, EVE segments replicated "actual" technology/asset capability. In Phase II, the next level of validation of the performance capabilities of products under test or within the test environment is established to support "firsthand" refinement of the EVE segments to ensure they represented "realism." With both complexity and realism validated, along with the CRA's understanding of intel quality effects of decisions, the CRA is now ready to develop battle readiness, engagement, and management support in the form of recommendations that have statistical significance.

Using the Pareto chart analysis approach, Figure 2, the CRA identifies a single EVE tree solution, again combining EVE segments using an AI/ML algorithm trained earlier, that supports as many point variations in the green zone as possible.

> Pareto front is a set of nondominated solutions, being chosen as optimal, if no objective can be improved without sacrificing at least one other objective. On the other hand, a solution x* is referred to as dominated by another solution x if, and only if, x is equally good or better than x* with respect to all objectives. (www.igi-global.com, n.d.)

The green zone is defined with user thresholds for probability of mission success from the Monte Carlo simulations. This is the optimal solution given a wider variety of influencer actions. The EVE tree representing these group of points in the green zone of the Pareto Chart is what is recommended to minimize effects of influencer variations and EVE tree weak points. There are two types of recommendations provided:

- Recommendation Type 1. Nominal EVE tree solution that includes as many points in the green zone as possible.
- Recommendation Type 2. Resilient EVE tree solution that supports ability to withstand a state variable flipped to zero but still support a successful mission. This resilient EVE solution must also be able to include as many points in the green zone as possible. This second type assumes unknown-unknowns occur, and because it was a surprise, successfully flipped a bit. The recommendation ensures continued success because of its solution resilience.

The "why" is provided to support explainable AI, not just in the causal relationships but also how those causal relationship caused statistical results, as collected in previous stages. The outcomes (plotted points) are known wargaming results, with variations of intelligence, depicting the Red Force's ability regarding what they do (state variable flipped), how they do it (EVE tree), and when the Blue Force is attacked. Thresholds, from one color region on the Pareto chart to the next, are determined by values calculated using the discrete correlation approach. Green area indicates that the value was considered a successful mission result when adjudicated, i.e., being above the threshold of what is considered a successful mission. Yellow areas indicate large variations regarding success and failure associated with the threshold that could not be resolved, therefore having outcomes that are uncertain. Upper, right yellow region indicates bias towards

Red Force likelihood of success. Lower, left yellow region indicated bias towards Blue Force likelihood of success.

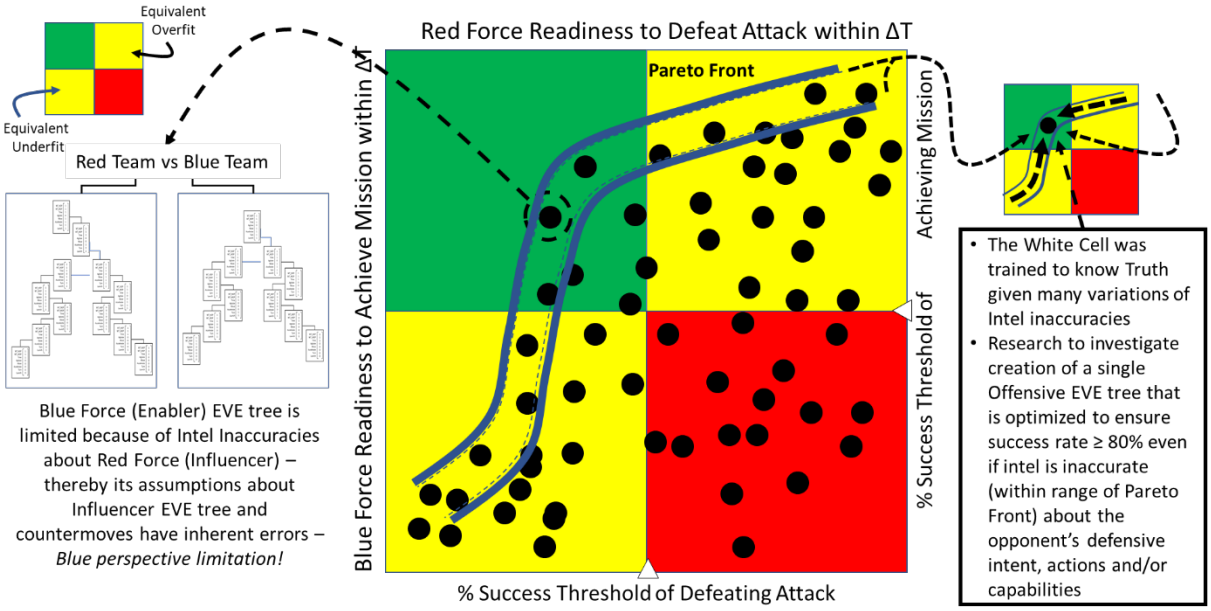

Figure 2. EVE Tree That Optimizes Ability to Succeed Independent of Opponent Strategy

As stated earlier, an "offense" action can include defensive tactics described in the EVE tree (Nagy, 2021; Nagy, 2022), and defensive actions can include offensive tactics, again described in the corresponding EVE tree. Consider how machine learning systems generalize. The CRA takes in instances of training data (points on the plot) to learn, through feedback, how to correctly determine the meaning of the input. The result is that the CRA is designed to handle variations from the original training data and still determine the meaning.

The "What's In It For Me" (WIFM) Human Factor

If users are to participate with using the CRA, there needs to be a significant return on investment for wargamers and operational test engineers, or why would they be motivated to change or do something different?

To support the motivation of professional wargamers, the goal is to automate their existing tool suite. This means moves and countermoves can be more easily entered and analyzed with significantly greater statistical precision. Likewise, since CRA following NSP can repeat the entire wargame in detail, action by action using EVE chain events, it can play out the statistical variations for "what if" analysis. It can determine confidence factors by EVE segment, meaning move and countermoves can be closely examined, even analyzing how quality levels of intel can impact strategy and tactics. Through its automation, CRA can reduce the time to set up and implement that wargame, focusing more on content and less on administration. It can create an ontology that allows for wargamers to more easily share perceptions and joint actions.

To support the motivation of operational test engineers, the CRA, following the NSP, can achieve near- and long-term goals associated with more effective analysis and testing. It can support near-term goals defined in the National Security Commission of Artificial Intelligence Final Report 2021 by providing automated decision support for constrained test scenarios that are challenged with creating "realistic" battle engagement test scripts and real/synthetic environments for autonomous systems, including manned and unmanned teaming. The CRA

can be used to standardize existing TEVV/LVC facilities by providing a cost effective, common simulation environment. It will provide more accurate analysis of the strengths and weaknesses of the product under test, support resilience, and provide statistically explainable test script scenarios.

As a long-term goal defined in the National Security Commission of Artificial Intelligence Final Report 2021, the CRA will eventually be able to test an autonomous system, or a system of autonomous systems, designed to dynamically learn and adapt during a manned and unmanned teaming operation. The CRA will provide real-time decision support and course of action recommendations and auto-generated scripts. This will allow TEVV/LVC facilities to accurately replicate synthetic environments requiring open-world simulations to adequately test adaptable, autonomous platforms required to perform a wide range of joint and coalition enhanced missions.

Conclusions

The paper recommends that for a CRA to gain trust from its human users, it must be designed to fill two gaps in its training and evolvement process before being operationally deployed:

(1) Wargaming Gap (1): The CRA must learn how to provide successful recommendations during wargaming that involves complex battle scenarios that include unanticipated, “out of the box” surprises by the opposing force, or even when poor intel quality or ability to receive accurate status from its own assets are experienced.
(2) Operational Testing Gap (2): The CRA must learn how to create test scripts in support of VVTE/LVC facilities by providing requirement coverage, but also create tests that help in analyzing performance using complex battle scenarios, that include unanticipated, “out of the box” surprises by the opposing force, or even when poor intel quality or ability to receive accurate status from its own assets are experienced.

If these two specific gaps were generalized as an architype for AI development, it would be to: (1) Have the AI learn from Subject Matter Experts (SMEs), where its learning can be continually tested/validated, thereby proving performance and (2) have the AI be involved with “real” technology, learning from firsthand experience what systems can and cannot do, where its learning can be continually tested/validated, thereby proving performance. A final key aspect to using this archetype is ensuring that any human involved with the training of the AI receive value, i.e., his or her WIFM factor is also filled during the process.

These are key aspects related to both critical learning gaps (DSB, 2009) that must be filled in any CRA before being deployed to ensure trust is earned. These critical gaps are addressed by the NSP using EVE chains, and must be filled to adequately prepare the CRA for “realistic” experiences during operational deployment. During this training process, the CRA must demonstrate learned knowledge to wargamers and operational test engineers in worst case conditions, as described.

Both training gaps (1) and (2) indicate a need to work with wargaming and operational test engineers to produce battle scenarios that can help anticipate the unexpected and design an optimal response into the training data. This training must include resiliency plans for when the surprise encounter by the opponent is successful. Human oversight is still a necessity for a CRA when unwanted loss of life or property is in jeopardy, but by filling these two gaps when designing the CRA, trust will be earned through reliable performance, demonstrating the ability to deal with size and complexity of a combat situation.

Following the NSP, the CRA is designed to motivate three types of customers for continued use of the AI product. For professional wargaming, the CRA automates part of the arduous appraisal process, and provides improved analytical results, that includes causal factors. It can uniquely support wargamers with three analyses of the wargame red and blue based on intel received on each opposing side, and white knowing "truth." The CRA will be able to reenact entire wargames to statistically analyze strategies and tactics, showing bottlenecks, strengths, and weaknesses, as well as needs to improve resiliency. It can alter the intel, simulate the entire war game again, and show what ifs, trends, and variations. The AI system can learn and share those statistical results regarding how to prepare better for unanticipated, "out-of-the-box" surprises in battle from a blue perspective.

For testers, this NSP-created CRA provides test threads that enable evaluators to consider all possible uses of a particular technology in anticipated and unanticipated, but possible scenarios. It could share the analytical and statistical knowledge gained through wargaming to support the operational test engineers in developing more tactically and strategically "realistic" test scripts. The CRA would provide an automation capability to reduce time and effort in developing test scripts and ensure adequate coverage of requirements.

For operators, the CRA becomes a trustworthy tool, enabling auto generation and comparison of viable COAs, with causal, explainable factors using EVEs. It can provide COAs that can minimize red effects when limited intel is available. Further, the CRA may infer red intent and identify possible unknown unknowns, which can reduce the number of tactical surprises blue might face.

The CRA, following the nine-stage approach, has the ability to deliver new ideas on what red might do dramatically increase the blue planner and decision-maker's mental models on possible future outcomes. Moreover, the modeling of the EVE chains and related recall of EVE segments enable very rapid re-planning and generate new ways to think about and achieve operational resilience. Explaining a recommendation or action is different from developing a relationship of trust that the recommendation or action will achieve the desired result. This paper concludes that if there is a desire to minimize human involvement in complex battle scenarios (for example to improve reaction time or avoid human loss), then the AI must be able to handle the unexpected. This means the AI must be trained to handle the unexpected.

Again, the NSP is based on achieving trust through relationship with a CRA well before operational deployment. This ability to earn trust from reliable performance with wargamers and testers fills the two gaps. Additionally, the need to introduce unanticipated/surprises associated with state variables during wargaming and operational testing will enhance the ability of U.S. armed forces to prepare and overcome, resulting in less fatality, operational cost, and escalation. (DSB 2015) By using EVE chains with the three phases containing nine segments, integrating theory and practicality, it presents the potential of changing the outcome of future conflicts through COA recommendations that optimally counter unanticipated, out-of-the-box surprises by the opponent and handle complex scenarios.

References

Adamski, M., & Pence, S. (2019). Thriving in uncertainty from predictive-to-probability-based assessments. *Military Review,* March–April.

Canadian Army Intelligence Regiment, Land Force Intelligence Centre. (2021). Combined arms lessons, learned from the Nagorno-Karabakh War. Canadian Department of National Defense.

Cancian, M. (2018). ~~*Avoiding*~~ *Coping with surprise in great power conflicts.* Center for Strategic and International Studies. https://www.csis.org/analysis/coping-surprise-great-power-conflicts

Cooter, R. D. (2000). Three effects of social norms on law: Expression, deterrence, and internalization. *Oregon Law Review, 79*, 1–22.

Defense Science Board. (2009). Capability Surprise, Volume I: Main Report. DSB Capability Surprise. Vol I. https://dsb.cto.mil/reports/2000s/ADA506396.pdf
Defense Science Board. (2009). Capability Surprise, Volume II: Supporting Papers. Vol II. https://dsb.cto.mil/reports/2010s/ADA513074.pdf
DSB Strategic Surprise Final Board. (2015). DSB Summer Study Report on Strategic Surprise. https://dsb.cto.mil/reports/2010s/2014_DSB_Strategic_Surprise.pdf
Eisenhower, D. D. (2022). Quotes. Presidential Library, Museum and Boyhood Home. https://www.eisenhowerlibrary.gov/eisenhowers/quotes
Frank, A. (2015). Complexity, psychology, and modern war. *Small Wars Journal.*
Hicks, K. (2021). Implementing responsible artificial intelligence in the Department of Defense [Memorandum]. https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/implementing-responsible-artificial-intelligence-in-the-department-of-defense.pdf
Miller, S., & Nagy, B. (2021). Interdependence analysis for artificial intelligence system safety. *Proceedings of the 18th Annual Acquisition Research Symposium.*
Nagy, B. (2021). Functional hazard analysis and subsystem hazard analysis of artificial intelligence/machine learning functions within a sandbox program. *Proceedings of the 18th Annual Acquisition Research Symposium.*
Nagy, B. (2021). *Tips for applying artificial intelligence to battle complexity* [Presentation]. Naval Applications for Machine Learning Symposium, San Diego, CA.
Nagy, B. (2021). *Using event-verb-event (EVE) constructs to train algorithms to recommend a complex mix of tactical actions that can be statistically analyzed* [Presentation]. Naval Applications for Machine Learning Symposium, San Diego, CA.
Nagy, B. (2022). *Battle Readiness Engagement Management (BREM) prototype: A wargaming tool to analyze "realistic," complex kill chains* [Presentation]. National Fire Control Symposium.
Nagy, B. (2022). *Deploying meta-models in warfighter laptops to drive "real-time," on-site course-of-action recommendations that solve complex kill chain solutions* [Presentation]. National Fire Control Symposium
Nagy, B. (2022). *Event-verb-event (EVE) constructs to allow machine learned systems to solve complex kill chain problems* [Presentation]. National Fire Control Symposium.
Nagy, B. (2022). *Fourteen tips to increase confidence in the performance of artificial intelligence (AI)/machine learning (ML) functions during the five stages of development* [Presentation]. National Fire Control Symposium.
Nagy, B. (2022). *Using game theory and machine learning to provide white cell automation support in the adjudication of war game* [Paper presentation]. SPIE Conference on Defense and Commercial Sensing.
Nimitz, C. (1965). Chester W. Nimitz letter to Charles L. Melson. Naval Historical Collection Archives. https://usnwcarchives.org/repositories/2/archival_objects/39995
Perla, P. (1987). War games, analyses, and exercises. *Naval War College Review.* https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=4335&context=nwc-review
Rielage, D. (2017). War gaming must get red right. *U.S. Naval Institute Proceedings, 143/1/1,367*. https://www.usni.org/magazines/proceedings/2017/january/war-gaming-must-get-red-right
Schmitt, J. (2008). Command and (out of) control: The military implications of complexity theory. http://www.dodccrp.org/html4/bibliography/comch09.html
Johnson, B. (2019). Artificial intelligence—An enabler of Naval Tactical Decision Superiority. *Association for the Advancement of Artificial Intelligence.* https://doi.org/10.1609/aimag.v40i1.2852
Logan, D. (2009). Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry. *Journal of Experimental Botany*, *60*(3), 712–714. https://doi.org/10.1093/jxb/erp043
Moses, L. B. (2007). Recurring dilemmas: The law's race to keep up with technological change. *Illinois Journal of Law, Technology, and Policy, 2*, 239–285.
Silva, A. (2013). Unit Five: Kamikaze Pilots and Shinto. *Cultural Anthropology.* https://anyasilva89.wordpress.com/2013/05/14/unit-five-kamikaze-pilots-and-shinto/

Acquisition Research Program
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net