Behavior Envelopes for Defining Performance Metrics in Complex Scenarios

Henry Phillips[1], Randolph Jones[2], Jeff Craighead[2], Michael Charlton[3],
CDR(S) Joseph Geeseman[4], Joseph Cohn[2], Lorraine Barghetti[5]

[1]Advanced Distributed Learning Initiative, Orlando FL,
[2]Soar Technology LLC, Orlando FL,
[3]2 Circle, Inc, Fallon NV,
[4]Naval Air Systems Command, Patuxent River MD,
[5]Air Force Research Laboratory, Dayton OH

henry.phillips.ctr@adlnet.gov, randy.jones@soartech.com, Jeff.craighead@soartech.com;
mikeusn2011@gmail.com;joseph.geeseman.mil@mail.mil; joseph.cohn@soartech.com;
lorraine.barghetti.civ@us.af.mil

ABSTRACT

Evaluation of decisions made in complex, multi-entity scenarios, such as those typical of LVC training, is extremely difficult (Carroll, 2023; Wang & Wang, 2024). In simple, low entity-count scenarios, it can be possible to rely on a library of goals and metrics defined for the evaluation of the performance of and decisions by a single operator, where there is one best option to select or correct decision to be made (Mbogu & Nicholson, 2024). As mission complexity and entity count increase, however, these assumptions fail. A critical need exists for informed evaluation of decisions made in complex situations in which it is not possible to unambiguously prioritize or achieve a set of entity-level goals and expectations (Tato, Nkambou, & Tato, 2023).

An alternative available for construction of complex metrics is the specification of behavior envelopes (Jones et al., 2015), a tool that can be used to define the set of conditions within which a system can function reliable and predictably (Wray et al., 2021). A Behavior Envelope consists of three primary components: (1) a situational context defining observable features of a situation, as well as unobservable features describing the internal state of the target being evaluated; (2) expectation constraints that the behavior should meet in situations where the behavior context applies; and (3) finally, a non-binary scoring/fit function that evaluates how well the behavior being observed conforms to the expectation constraints.

This paper explores the application of behavior envelopes to multi-entity performance and decision evaluation in complex scenarios, through analysis of a corpus of non-deterministic, high iteration count synthetic, labeled Semi-Autonomous Forces (SAF) defensive counter-air (DCA) mission iterations. Specification of the context, expectations, and scoring/fit of the observed complex mission envelopes provide the basis for specification and quantification of the achievement of multi-entity mission criteria.

ABOUT THE AUTHORS

Henry L. Phillips IV, PhD is Program Manager for the Advanced Distributed Learning (ADL) Initiative. He served 20 years on active duty as a Naval Aerospace Experimental Psychologist including service as Executive Officer of the Naval Air Warfare Center Training Systems Division. His published work and presentations focus on training, selection, and simulation.

Randolph M. Jones, PhD is a Senior Artificial Intelligence Engineer and co-founder of SoarTech. He has 40 years of experience doing research and development in artificial intelligence and machine learning, much of it focused on AI solutions for DoD training and simulation. He earned his PhD in Information and Computer Science in 1989 from the University of California, Irvine.

Jeff Craighead, Ph.D. is a Lead Scientist at Soar Technology, LLC. His work focuses on cognitive AI and behavior-based architectures for robotics, simulation, and game-based training. Jeff has extensive experience in developing and commercializing desktop, mobile, and embedded applications.

Michael Charlton is an instructor and analyst with 2 Circle Consulting, Inc. He also serves as an E/A-18G Weapons Tactics Instructor at NAS Fallon and is trained as a Naval Flight Officer. He holds a BS in Astrophysics from the United States Naval Academy and is completing his MA in data science from UC Berkeley.

LCDR Joseph Geeseman, PhD is the Military Director for the LVC program of record at the Naval Aviation Training Systems and Ranges Program Office (PMA-205). He also provides strategic planning for PMA-205 and manages the extensive research and development portfolio for the program office.

Joseph Cohn, PhD, Director of SoarTech's Readiness and Medical Solutions team, is a retired Navy Medical Service Corps Captain whose career has focused on high-risk research informed by requirements to deliver solutions that ensure the United States maintains its technical edge over its adversaries. Joseph has proven expertise envisioning and advancing biomedical and human-machine interface solutions informed by emerging technologies, like Artificial Intelligence, brain-machine interfaces and wearable sensors, supporting Human System, Medical, C4ISR, and Manned-Unmanned Teaming applications. Joseph is an Associate Fellow of the Aerospace Medical Association, a Fellow of the Society for Military Psychology and the American Psychological Association.

Lorraine Borghetti, PhD, is a cognitive scientist and research lead at the 711 Performance Wing, Air Force Research Laboratory at Wright Patterson Air Force Base, Ohio. Her research focuses on the role of contextuality and sequential effects in cognition and learning at individual as well as human-machine team level.

Behavior Envelopes for Defining Performance Metrics in Complex Scenarios

Henry Phillips[1], Randall Jones[2], Jeff Craighead[2], Michael Charlton[3],
LCDR Joseph Geeseman[4], Joseph Cohn[2], Lorraine Barghetti[5]

[1]Advanced Distributed Learning Initiative, Orlando FL,
[2]Soar Technology LLC, Orlando FL,
[3]2 Circle, Inc, Fallon NV,
[4]Naval Air Systems Command, Patuxent River MD,
[5]Air Force Research Laboratory, Dayton OH

henry.phillips.ctr@adlnet.gov, randy.jones@soartech.com, Jeff.craighead@soartech.com;
mikeusn2011@gmail.com;joseph.geeseman.mil@mail.mil; joseph.cohn@soartech.com;
lorraine.barghetti.civ@us.af.mil

THE PROBLEM OF EVALUATING COMPLEX SYSTEMS OF BEHAVIOR

One of the most critical problems facing US military aviation may be a need for a deeper understanding of the factors and metrics that can predict, and therefore help improve, mission outcomes for complex multi-entity events. The next conflict the US enters is highly likely to involve complex scenarios involving high entity counts, multiple combat domains, and extremely broad battlespace. Understanding the factors underlying individual and team performance and mission success or failure is critically needed. Literature has not kept pace with the demand for this level of analysis, unfortunately. Available literature tends to focus on the prediction of the behaviors of aircraft in narrowly defined contexts involving highly standardized circumstances and parameters (Li et al, 2022; Tato et al., 2023; Zeng et al., 2020)., or on non-algorithmic qualitative summaries of the factors affecting complex mission outcomes (Carroll, 2023).

This work is intended to add to those works that strike a balance between these two extremes: Wang and Wang (2022) looked at decision-making among autonomous combatants in multi-entity sorties, which is certainly relevant, though these entity decision processes will differ significantly from those made by human pilots for many reasons. Mbogu and Nicholson (2024) have modeled root cause analyses for multi-entity mission outcomes by comparing graphs of expected to observed behaviors, which has significant relevance to this problem space.

In this vein, the purpose of this work is to provide guidelines and examples for quantitative, modellable performance indices at the individual entity and group levels in complex events, borrowing from Wray and colleagues (2021) and Jones and colleagues (2015). The behavior envelope modeling technique described here is a) objective and algorithmic, b) can be applied to complex datasets in which only a limited subset of possible factors and constraints may be known or modellable, and c) can be used to small n data stacks for which only a limited number of iterations are available. This approach relies on SME-derived context cues to attach meaning to observed combinations of events within a larger scenario context. These events can be used to attach meaning and extract information from larger fields of observed events and behaviors. We can use these envelopes to help build the conditions to recognize the relevant behavioral *needles* in what can be an enormous *haystack* of potentially relevant events comprising a complex multi-entity mission.

Evaluation of intelligent decision-making behavior remains an open problem (Wallace, 2003). Even the evaluation of formal and ostensibly deterministic software systems becomes intractable as the complexity of their expected operations increases. When the requirement that informal systems (such as humans) are expected to be “smart” and possibly even “innovative”, this intractability increases by orders of magnitude. This complexity can increase by another order of magnitude when we move from evaluating the performance of individual decision makers to evaluating the aggregate behaviors and outcomes of groups of decision makers. Increases in complexities are driven primarily by the facts that measures of success can come in degrees, are multi-factorial, and are sensitive to context and nuance.

A significant example of this complexity and difficult is the evaluation of decisions made in complex, multi-entity scenarios, such as those typical of LVC training. In simple, low entity-count scenarios, it can be possible to rely on a library of goals and metrics defined for the evaluation of the performance of and decisions by a single operator, where there is a single best option or correct decision. For example, subjective assessments by subject-matter-experts (SMEs) can produce effective and consistent evaluations. However, as mission complexity and entity count increase, the assumptions of "best options" and "correct decisions" fail. This in turn leads to subjectivity and inconsistency in assessments. A critical need exists for informed and nuanced, but objective, evaluation of decisions made in complex situations in which it is not possible to unambiguously prioritize or achieve a set of entity-level goals and expectations. We need a way to evaluate what happens in situations too complex for a static evaluation rubric.

Jones et al. (2015) have previously described and applied the concept of ***behavior envelopes***, a formalism that can be used to define the window of variation around which a set of goals or behaviors can be interpreted to have been met or to have occurred. That is, behavioral envelopes represent the set of conditions within which a system can function reliably and predictably. A Behavior Envelope consists of three primary components: A ***situational context*** defining observable features of a situation, as well as unobservable features describing the internal state of the target being evaluated; ***expectation constraints*** that the behavior should meet in situations where the behavior context applies; and finally, a non-binary ***scoring/fit function*** that evaluates how well the behavior being observed conforms to the expectation constraints.

Behavior envelopes have seen several applications, including planning, intent recognition, adversarial reasoning, and behavior prediction. The roots of behavior envelopes, however, are in assessment, and have typically been used to evaluate individual performance of humans and synthetic forces (Jones et al., 1999; Wallace & Laird, 2003). This paper explores the extension of behavior envelopes to the use case of multi-entity performance and decision evaluation in complex scenarios, through analysis of a corpus of non-deterministic, high iteration count synthetic, labeled Semi-Automated Forces (SAF) mission iterations. Here, behavior envelopes serve as the basis for understanding the causes of collective mission success or failure, beyond what can be yielded through analysis of individual performance or formalized Techniques, Tactics, and Procedures (TTP) adherence.

In addition to describing or predicting mission success or failure, behavior envelopes can help explain the causal antecedents of success or failure in complex events – a requirement that to date is met exclusively by SME analyses, which is not only time and labor intensive, but is also subjective, thus producing potentially varying results across SMEs with different evaluation priorities.

INSPIRATION FROM METHODS FOR SOFTWARE VERIFICATION AND VALIDATION

Behavior envelopes are a generalization of software system verification and validation methods. Formal verification and validation methods provide evidence that systems perform as expected under all conditions evaluated. Formal methods quickly become intractable as software complexity scales up, for two primary reasons:

- It becomes increasingly difficult to define a set of requirements that describe all aspects of the system's required behavior.
- It becomes increasingly difficult to anticipate all the different situations in which things could potentially go wrong.

In the area of software engineering, a pragmatic response to these issues has been the development of "property-oriented specifications" (Dasso & Funes, 2009). Property-oriented specification languages facilitate automated validation of software systems by asserting specific relationships between elements of a system's data or behavior. An example property-oriented assertion is "At this point in the program's sequential logic, the value of the velocity variable must be greater than zero." A significant pragmatic advantage is that property-oriented specification does not need to be complete to be useful or to be implemented. Automated software can monitor various assertions about properties and report any violations. Engineers can put their energy into defining through specifications only in the most critical areas of the software. Property-oriented specifications therefore have some appeal for application to the validation of intelligent system behavior. However, there are two issues to be resolved in adopting a similar approach to validating complex intelligent behavior.

First, the *context* of a specification must be determined. In standard software systems, assertions about behavior occur at the sequential execution point in the code at which those assertions are applicable. This is feasible because, even for complex software, there are individual threads of execution that define the “sequential location” of the execution logic at any point in time. In contrast, intelligent behavior involves much more loosely bound goals, methods for achieving the goals, and processes for making sense of the world. In an intelligent system or in a person performing a complex task, all these processes must interleave flexibly in ways that make it difficult to recognize a “state” for the system. We desire the ability to specify the context in which some property holds, including contexts that may not be entirely observable.

Second, intelligent behavior is rarely usefully classified as simply “correct” or “incorrect.” Intelligent behavior is varied and flexible, and competence for accomplishing goals occurs in varying degrees. In addition, competing goals play out within a trade-space, with different losses and gains dependent upon context and consequences. Thus, we desire a specification language that supports validation scoring functions that are not simply binary. The automated specification system should be able to indicate the degree to which an observed behavior meets (or fails to meet) the specification, rather than simply reporting that it fails to meet the specification.

Behavior envelopes take advantage of the strengths of property-oriented specifications while also addressing the challenges identified above. A primary advantage of behavior envelopes is that they allow the specification of evaluation criteria to an arbitrary level of detail. As with property-oriented specifications for traditional software systems, this allows users to create inexpensive but useful behavior specifications, or to invest in more detailed specifications for high-priority behaviors that justify the additional investment. The consequence is that behavior envelopes scale to the evaluation of large, complex systems by supporting focus of effort and detail in the most critical areas of interest.

SCALING IN COMPLEXITY AND SIZE

Behavior envelopes provide a useful tool for evaluating complex behavior. They make easier, but do not replace, the hard work of defining the *requirements* and *metrics* to be evaluated. Those must come from SMEs, data analysis, or other processes. A significant advantage of behavior envelopes is that they can be specified as precisely as they need to be, and no further (Jones et al., 2015). It is possible to define requirements and scoring functions that are not 100% comprehensive, but that are still highly useful: Behavior envelopes thus enable a pragmatic approach to automated evaluation.

In addition, in cases where a particular behavior envelope is found to be insufficiently detailed, SMEs can focus their work on increasing level of detail. The advantages of behavior envelopes include that *they do not need to be complete to be useful*, and the requirements-definition work required to improve them can be focused only on the areas in which the additional effort pays off in the form of *improved evaluations*.

In addition to scaling in terms of complexity, behavior envelopes support scaling in terms of the size of the targets being evaluated. Because behavior envelopes are configured to boundaries of a situational context, rather than individual points in execution sequences, the envelope contexts can just as easily describe requirements on *group behaviors* as on individual behaviors. All that is necessary to apply behavior envelopes to groups is to define the group contexts to which a particular evaluation metric should apply, and then construct the aggregate expectations and scoring functions that are appropriate to that group (Li et al., 2022). The remainder of this paper introduces and analyses some applied use cases, demonstrating the extension and scaling of behavior envelopes to the evaluation of group behavior.

A NAVAL AVIATION USE CASE

In the context of naval aviation, behavior envelopes can be used for a variety of purposes, such as contributing to the well-established corpus of individual performance measures. Alternatively, a unique and novel application of behavior envelopes would be to characterize when and where highly contextualized events occur --involving *many entities* – as a method to reveal the underpinnings of set of mission outcomes not otherwise readily discernable.

For single operator, decision maker, or asset performance evaluations, behavior envelopes can be used to recognize and characterize events of importance (Wray et al., 2021). Moreover, they can be used to define and recognize violations of TTP, mission briefing constraints, or deviations from an a priori plan. Here, we explore extending their utility in characterizing multi-entity events taking place during complex missions. As the number of assets involved in a mission increases, so does the probability something can go awry, even without overt violations of TTP by any single asset. Applying behavior envelopes to multi-entity behavior captures three distinct categories of information: a) the increased complexity of **context** describing mission setting, conditions, and environment, as well as considerations regarding how involved entity decisions may change that environment; b) **expectations** regarding what will happen as time progresses as multiple decisions are made during the mission; and c) **evaluation** of those decisions and resulting contextual changes via a **scoring function** to help clarify why conditions changed, and ultimately why a mission outcome happened.

Behavior envelopes can be informative for SMEs/Evaluators watching complex operational or training events, and foster consistency and objectivity of evaluations across groups of evaluators with varying backgrounds and priorities. One of the critical tasks in building informed debriefs/AAR is the development of root cause analyses (RCA) of observed mission outcomes. Building these RCAs requires SMEs or systems to implicitly use and rely on the information captured in multi-entity behavioral envelopes. A debrief system that captures and annotates such occurrences can automate a portion of the RCA process for SMEs, reducing the time and workload required to generate these insights. Under AFRL funding, we are separately pursuing collaborative tools that allow analysts to extract patterns from existing data sets and search for potential RCAs. While other tools already exist that can predict mission success or failure (e.g., kill ratios, asset survival), multi-entity complex mission behavior envelopes can help SMEs/stakeholders more quickly understand *why* a mission succeeded or failed in its objectives.

Consider the following example, in which a mission is conducted by a set of 4 blue kinetic entities against 3 red surface to air missile (SA) sites and a ground target, as depicted in the upper half of Figure 1. We use this example to define a set of behavior envelopes intended to capture details relevant to the evaluation of how well Blue 1-4 manage their time-fuel-weapons (TFW) resources, both *individually* and *collectively*.

Time-Fuel-Weapons Behavior Envelope for Capturing Individual Decisions

Let us consider the case in which we evaluate the ***individual decisions*** made by the aviators controlling the blue entities separately. We can define initial contextual cues as follows:

Context

- The mission plan as briefed has accounted for all known threats (in this instance, a total of 3 SA sites) and ingress route.
- The mission plan includes a partition of targets that allocate one to each blue asset: Blue 1 – 3 are each tasked with eliminating one of the SA sites, while Blue 4 is expected to eliminate the mission target with a strike weapon.
- We can add a third contextual element that stipulates the threats as briefed are in fact the threats that will be encountered (this becomes relevant later).

Expectations

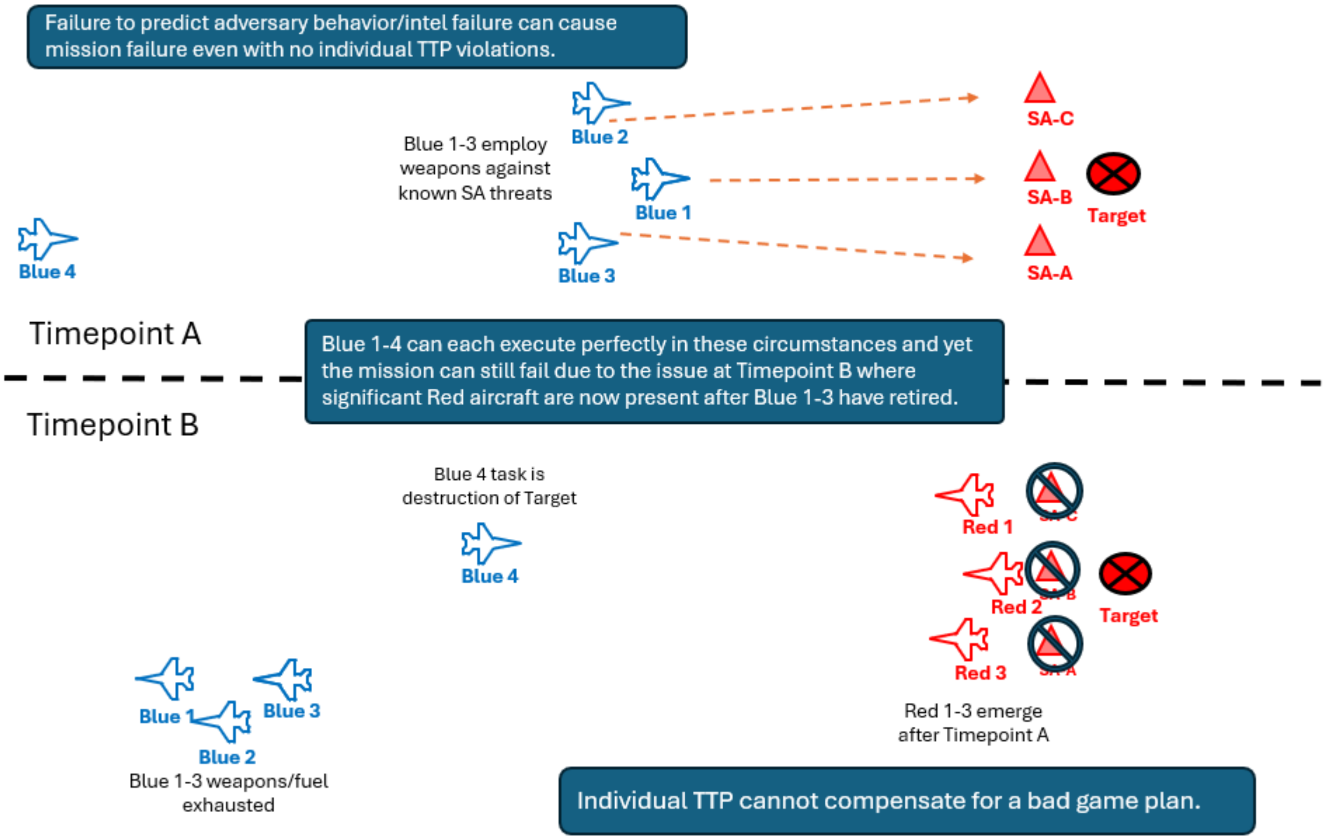
These context definitions dictate the situation bounds in which there are several expectations:

- Assets Blue 1 – 3 expend fuel and weapons according to the mission plan, in a manner necessary to accomplish all objectives.
- Assets Blue 1 – 4 are likely to accomplish their objectives, with Blue 1-3 eliminating the SA sites and Blue 4 destroying the ground target.

Scoring

These context definitions and expectations in turn can be evaluated using several possible scoring functions:

- The decisions by blue 1-3 to continue the mission up to and beyond their ordinance release points were correct, and should be assigned some positive value.
- The decisions by blue 1-3 to egress once weapons were expended according to the planned route were correct.
- Blue 1-3 success in destroying SA sites should be assigned some positive value, and Blue 4 success in destroying the mission target should be assigned some positive value as well.


Failure to predict adversary behavior/intel failure can cause mission failure even with no individual TTP violations.

Blue 1-4 can each execute perfectly in these circumstances and yet the mission can still fail due to the issue at Timepoint B where significant Red aircraft are now present after Blue 1-3 have retired.

Individual TTP cannot compensate for a bad game plan.

Blue 1-3 employ weapons against known SA threats
Blue 2
Blue 1
Blue 3
Blue 4
SA-C
SA-B
SA-A
Target
Timepoint A
Timepoint B
Blue 4 task is destruction of Target
Blue 4
Red 1
Red 2
Red 3
Target
Red 1-3 emerge after Timepoint A
Blue 1
Blue 3
Blue 2
Blue 1-3 weapons/fuel exhausted

Figure 1. A Time-Fuel-Weapons (TFW) Management Failure Scenario Example.

We can additionally define a simple, multi-entity behavior envelope to reflect the degree to which these assets ***manage their time, fuel, and weapon (TFW) resources*** based on a set of contextual cues, expectations, and possible scoring functions. This simple envelope will become far more complex as we consider other possible permutations to be accounted for. Suppose, as is depicted in the lower half of Figure 1, that three Red assets (Red 1-3) not included or planned for in the mission briefing arrive in the mission space immediately following the retirement of Blue 1 -3. Circumstances will change substantially, and at this point, options available to blue assets may be limited or eliminated. In such cases, we still require quantitative and context-sensitive means to evaluate outcomes and alternatives.

Additional Context

- Additional threats emerge during the mission after Blue 1-3 make the commitment to engage their planned targets.
- Blue 4 now faces a superior red force alone with no escorts.

Expanded Expectations

- Blue 1-3 may continue on their originally briefed course.
- Blue 1-3 may neutralize their 3 targeted SA threats.
- Blue 4 may choose to retire from the engagement without destroying the target.
- Blue 4 may engage the red force and be destroyed. (Let us assume for purposes of this example that additional outcomes in which Blue 4 manages to destroy the target and then either escapes or is destroyed are essentially impossible.)

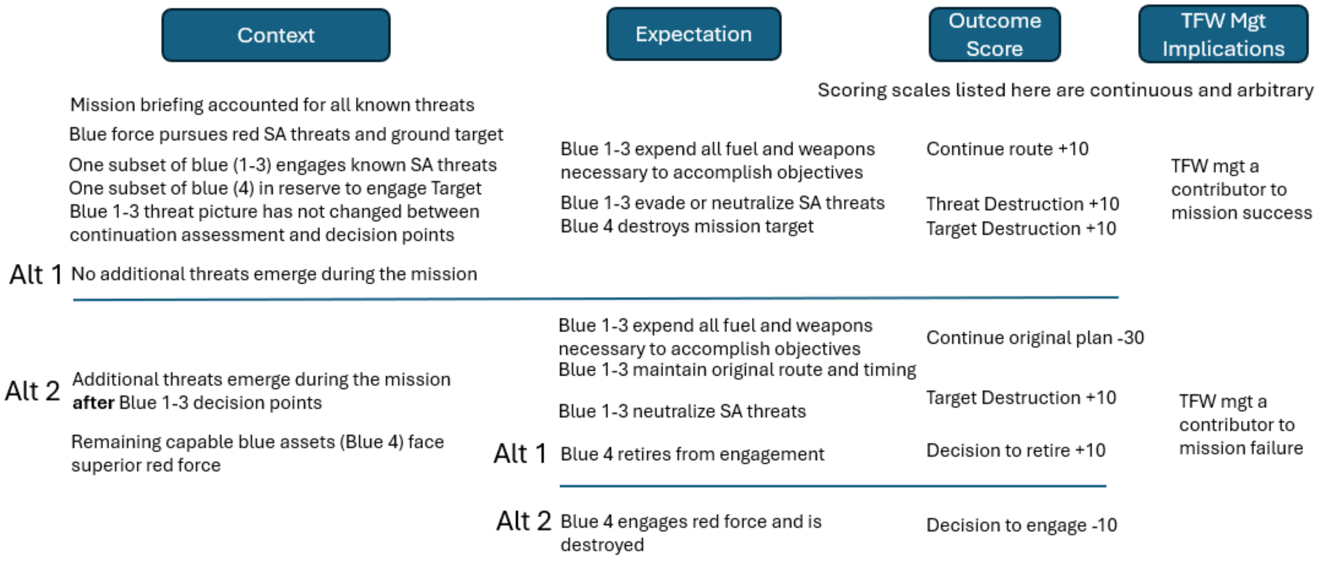
Scoring

Using a continuous scale to assign scoring values to all the above options is an excellent idea, particularly since none of the options seem particularly attractive, nor definitively correct. Some are worse than others, but none are "ideal."

- The option for Blue 1-3 to continue on their original plan, with no modifications to protect Blue 4, can be assigned a negative value, but is not definitively "incorrect."
- Possible elimination of assigned targets by Blue 1-3 are still mission objectives and will contribute to the possibility of survival by Blue 4, as well as the likelihood of mission success, and should be assigned positive values.
- The option for Blue 4 to retire from the engagement increases the likelihood that Blue 4 will survive but eliminates the possibility of mission objective achievement. Nonetheless, this may be assigned some positive score based on the value of an outcome that preserves the strike assets.
- The option for Blue 4 to engage the emergent red forces and be destroyed is assigned a small negative score in this example, under the assumption that a blue entity facing overwhelming enemy force should have a realistic idea of whether such an engagement will be survivable.

The example as described so far is presented as a function, possibly a weighted sum of a set of decisions by the individual blue force pilots, all of which were assigned quantitative values. These values were assigned using scoring functions and defined for purposes of algorithmic recognition by a system in terms of their definition by cues of context and expectation. The functions themselves can be derived and calibrated by SMEs, together with analytical tools and simulations to confirm or fine-tune their choices.

See Figure 2 for a summary of the context, expectations, and scoring elements associated with this example, as well as additional permutations and extensions of it discussed immediately below.



Context

Expectation

Outcome Score

TFW Mgt Implications

Scoring scales listed here are continuous and arbitrary

Mission briefing accounted for all known threats
Blue force pursues red SA threats and ground target
One subset of blue (1-3) engages known SA threats
One subset of blue (4) in reserve to engage Target
Blue 1-3 threat picture has not changed between continuation assessment and decision points

Alt 1 No additional threats emerge during the mission

Blue 1-3 expend all fuel and weapons necessary to accomplish objectives

Blue 1-3 evade or neutralize SA threats
Blue 4 destroys mission target

Continue route +10

Threat Destruction +10
Target Destruction +10

TFW mgt a contributor to mission success

Alt 2 Additional threats emerge during the mission **after** Blue 1-3 decision points

Remaining capable blue assets (Blue 4) face superior red force

Blue 1-3 expend all fuel and weapons necessary to accomplish objectives
Blue 1-3 maintain original route and timing

Continue original plan -30

Blue 1-3 neutralize SA threats

Target Destruction +10

Alt 1 Blue 4 retires from engagement

Decision to retire +10

Alt 2 Blue 4 engages red force and is destroyed

Decision to engage -10

TFW mgt a contributor to mission failure

Figure 2. Context, Expectation, and Scoring Cues from the TFW Management Failure Scenario Example for evaluation of Individual Decisions.

The results produced by these scoring functions contribute, in part, to the bigger picture for mission stakeholders of how time-fuel- weapons management may have affected mission outcomes. An essential feature that must be observed is that, in the scenario above, *mission failure cannot be attributed to any single decision by any individual operator*. For purposes of building mission level metrics, and identifying causal contributors to mission outcomes, we must describe the mission in collective terms. Behavior envelopes give as a formalism for doing this.

Time-Fuel-Weapons Behavior Envelope for Capturing Collective Decisions

We next turn to the question of how we can evaluate the ***collective decisions*** made by the aviators representing the blue force, as depicted in Figure 3, below.

Scoring

These context definitions and expectations in turn can be evaluated using several possible scoring functions:

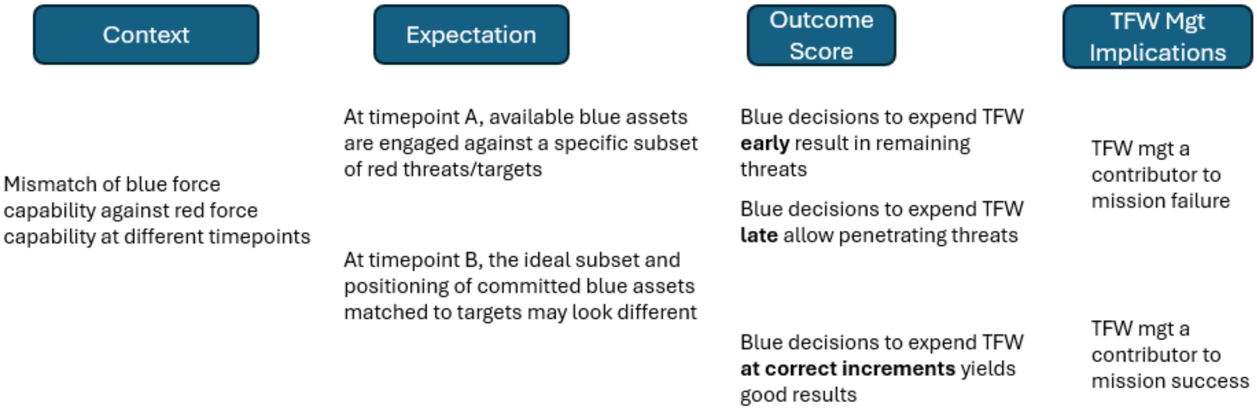
- The decisions by blue 1-3 to continue the mission up to and beyond their ordinance release points were correct, and should be assigned some positive value.
- The decisions by blue 1-3 to egress once weapons were expended according to the planned route were correct.
- Blue 1-3 success in destroying SA sites should be assigned some positive value, and Blue 4 success in destroying the mission target should be assigned some positive value as well.

Context

- *Repeated:* The mission plan as briefed has accounted for all known threats (in this instance, a total of 3 SA sites) and ingress route.
- *Repeated:* The mission plan includes a partition of targets that allocate one to each blue asset: Blue 1 – 3 are each tasked with eliminating one of the SA sites, while Blue 4 is expected to eliminate the mission target with a strike weapon.
- *Repeated:* Additional threats emerge during the mission after Blue 1-3 make the commitment to engage their planned targets.
- *New:* The relative balance of blue force capability to red force capability can be captured at different points in time. This can include the relative strengths and vulnerabilities of each of the blue assets to the red assets in the scenario, which is also a function of distance and positioning relative to other assets across timepoints.

Expectations

- At timepoint A, available blue assets are committed or engaged against a specific subset of red threats and targets.
- At subsequent timepoints, available blue assets may be committed against different threats and targets.
- Different combinations of target commitments and entity vectors can be assigned values defining their suitability and optimization for achievement of mission objectives.
- The optimal engagement and targeting solution at any given timepoint can be assigned a numeric value.
- These values will differ across timepoints based on the presence, location, and posture of red threats and blue assets.



Context

Expectation

Outcome Score

TFW Mgt Implications

Mismatch of blue force capability against red force capability at different timepoints

At timepoint A, available blue assets are engaged against a specific subset of red threats/targets

At timepoint B, the ideal subset and positioning of committed blue assets matched to targets may look different

Blue decisions to expend TFW **early** result in remaining threats

Blue decisions to expend TFW **late** allow penetrating threats

Blue decisions to expend TFW **at correct increments** yields good results

TFW mgt a contributor to mission failure

TFW mgt a contributor to mission success

Figure 3. Context, Expectation, and Scoring Cues from the Time-Fuel-Weapons (TFW) Management Failure Scenario Example for evaluation of Collective Decisions.

Scoring

For purposes of evaluating contributions of these expectations to specific mission outcomes:

- Blue decisions to expend TFW to target assets *earlier* than they should be, according to target optimization scoring (as is overwhelmingly likely in this case, with the impending arrival of Red 1-3, as depicted in Figure 3.) will result in threats remaining after exhaustion of blue forces.
- Similarly, blue decisions to expend TFW to target assets *earlier* than they should be expended will increase the likelihood that threats can penetrate the mission area beyond the barrier created by the blue assets.

- The distance in time between the expenditure of TFW as observed and the optimization maximum can be assigned a numeric value.
- This numeric value is a *direct indication of how well TFW management was executed* and can be used by SMEs in generation of RCA estimates, or directly as a metric relevant to mission effectiveness.

SUMMARY, CONCLUSIONS, AND FUTURE WORK

The example described above illustrates how it can be possible for individual operators to follow TTPs as prescribed and still experience mission failure. In the context of such unexpected outcomes, behavior envelopes help explain the circumstances contributing to mission success or failure. Extending this capability to the evaluation of multi-entity scenarios provides causal insights otherwise not easily detected by traditional evaluation methods. Weapons Tactics SMEs routinely conduct root cause analyses (RCA) as part of their after-action reviews to understand why complex missions experience success or failure, by evaluating all the factors affecting the battlespace and how they affect one another in causal chains leading to mission outcomes. Collective behavior envelopes can be used to help identify those causal factors, supporting the RCA work of Weapons Tactics SMEs, and contributing to meaningful, explainable metrics for the evaluation of why complex missions succeed or fail.

There are already numerous tools and models that adequately capture when and whether missions fail, and even tools that predict success or failure (i.e., objective achievement, kill ratios) for complex events. The gap to be addressed by the proposed Behavior-Envelope formalism is to produce a tool that will a) help Weapons Tactics SMEs understand *why* missions succeed or fail, and b) attach reliable, replicable quantitative values to those assessments.

Behavior Envelopes as the Basis for Complex Aviation Mission Effectiveness Metrics

Behavior envelopes have already been successfully used to assess individual performance metrics for aviation missions (Jones et al., 2015), and from prior projects we have developed an initial set of individual performance and behavior metrics for air intercepts. Exploiting the ability to scale behavior envelopes in terms of size and capability, we can use SMEs to expand this initial set to cover situations and metrics for mission effectiveness of various group sizes and different levels of aviation mission complexity. Once the behavior envelopes have been defined, they provide objective, automated computation of evaluation scores, thus reducing the manpower costs and subjectivity involved in using SME assessments and ratings.

Integrating Behavior Envelopes with Existing Capture of Individual Metrics

An additional advantage of behavior envelopes is that the scoring functions can themselves be composed from other scoring functions. Alternatively, they can use existing evaluation metrics “as is”, simply by specifying the contexts in which the existing metrics are to be used, or the desired metric values associated with different contexts. Thus, for example, if there are existing tools that already provide individual performance metrics to be visualized and assessed by SMEs, much of the work of creating the behavior envelopes is already complete. If the SMEs can define their context and threshold criteria for accepting or rejecting performance based on the existing metrics, these can be combined with the existing performance evaluation functions to create new envelopes.

Enhancing SME Analysis with Data Generation and Analysis

For the future development of reusable libraries of behavior envelopes, it will be essential to combine the expertise of SMEs with the generation and analysis of large amounts of data. Such a collaborative effort will be necessary both for developing a significant set of aviation-related behavior envelopes and for characterizing (and fine tuning) the quality of the behavior envelopes themselves. We have therefore additionally investigated methods for data generation and collaborative creation of behavior envelopes (Jones, Bechtel, & DeGrendal, 2018). One existing data generation and analysis tool allows an analyst to specify a scenario template plus a set of template parameters, together with value distributions for each parameter. The system then performs Monte Carlo sampling of the parameter value distributions to instantiate the scenario template to the Navy’s Next Generation Threat System (NGTS) simulations, producing numerous variations of the template. By running these varied scenario instances in NGTS, we can generate large numbers of datasets that cover the space of possible mission progressions and outcomes. SME analysts can then collaborate with a toolkit of data-analysis, machine-learning, and search-based tools, operating over these datasets to provide extensive summaries of possible progressions and patterns. These provide the elements that can be further classified based on outcomes to rate different categories of mission effectiveness. These classifications and categorizations will provide the basis of the context and scoring definitions for new behavior envelopes, as well as a gauge against which to measure the quality of behavior envelope assessments.

CONCLUSION

Previous work on behavior envelopes have described the computational mechanics of these tools (Wray et al., 2021; Jones et al., 2015). This paper expanded on those with a more specific aviation example, and direct linkage of mission context to the usefulness of the metrics to be described. The discussion explored a simple example of how the criteria for defining a behavior envelope could be used as the basis for defining a mission relevant set of metrics for explaining a single dimension of aviation performance evaluation, TFW management. More importantly, we explored how the evaluation of TFW management at the level of individual operator decisions was inadequate for understanding the role that this aspect of performance played in determining why and whether the mission succeeded. Finally, we discussed how the criteria used for mapping behavior envelopes could be applied to the evaluation of TFW management at the collective level.

The work to be done now is the specification of a library of mission relevant behavior envelopes for evaluation of individual and collective performance across dimensions, and generation of a data corpus to support a range of evaluations. This paper explained how the behavior envelope concept can be used to define the metrics needed to help tactical aviation SMEs better understand why success or failure was the outcome of a complex mission, and what decisions resulted in those outcomes.

ACKNOWLEDGEMENTS

The authors would like to convey their gratitude to Dr. Winston Bennett and Dr. Lorraine Borghetti for their guidance on the work upon which this effort was based, funded under Contract FA238423C0008 by the Air Force Research Laboratory. Distribution A: This work is approved by for public release; distribution is unlimited (AFRL-2024-3996).

REFERENCES

Carroll, M. (2023) Decision making in aviation. In Eds. Keebler, J. R., Lazzara, E. H., Wilson, K. A., & Blickensderfer, E. L., Human Factors in Aviation and Aerospace (Third Edition), Academic Press, 563-588. ISBN 9780124201392, https://doi.org/10.1016/B978-0-12-420139-2.00016-2.

Dasso, A., & Funes, A. (2009). *Formalization process in software development*. IRMA International.

Gupta, S., Davidson, J., Levine, S., Sukthankar, R., & Malik, J. (2017) Cognitive mapping and planning for visual navigation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2616–2625.

Jirgl, M., Havlikova, M., & Bradac, Z. (2015). The Dynamic Pilot Behavioral Models. *Procedia Engineering, 100.* DOI 10.1016/j.proeng.2015.01.483

Jones, R.E., Bachelor, B., Stacy, W., Colonna, Romano, J., & Wray, R.E. (2015). Automated monitoring and validation of synthetic intelligent behavior. *International Conference on Artificial Intelligence, 252-258, ICAI15*. (https://worldcomp-proceedings.com/proc/p2015/ICAI_contents.html)

Jones, R. M., Bechtel, R., & DeGrendel, B. G. (2018). Configurable adversary response prediction: Building efficient expectation models from high-fidelity behavior simulations. In *Proceedings of the 2018 Winter Simulation Innovation Workshop (SIW).* Orlando, FL.

Jones, R. M., Laird, J. E., Nielsen, P. E., Coulter, K. J., Kenny, P., & Koss, F. V. (1999). Automated intelligent pilots for combat flight simulation. *AI Magazine, 20*(1), 27–41.

Li, B., et al. (2022) A decision-making method for air combat maneuver based on hybrid deep learning network. *Chinese Journal of Electronics 31*, 107–115.

Li, B., Liang, S.Y., Tian, L.Y., et al. (2019) Intelligent aircraft maneuvering decision based on CNN. *International Conference on Computer Science and Application Engineering*, Sanya, article no.138.

Pellegrini, S., Ess, A., Schindler, K., & v. Gool, L. (2009) You'll never walk alone: Modeling social behavior for multi-target tracking. IEEE Int. Conf. on Computer Vision.

Tato, A., Nkambou, R., Tato, G. (2023). Automatic Learning of Piloting Behavior from Flight Data. In: Frasson, C., Mylonas, P., Troussas, C. (eds) Augmented Intelligence and Intelligent Tutoring Systems. ITS 2023. Lecture Notes in Computer Science, vol 13891. Springer, Cham. https://doi.org/10.1007/978-3-031-32883-1_48

Mbogu, H. M., & Nicholson, C. D. (2024). Data-driven root cause analysis via causal discovery using time-to-event data. Computers & Industrial Engineering, Volume 190, ISSN 0360-8352, https://doi.org/10.1016/j.cie.2024.109974.

Wallace, S. (2003). *Validating Complex Agent Behavior*, Ph.D. Thesis. The University of Michigan, Ann Arbor.

Wallace, S., & Laird, J. E. (2003). Behavior Bounding: Toward Effective Comparisons of Agents & Human Behavior, *International Joint Conference on Artificial Intelligence*.

Wang, H., Wang, J. Enhancing multi-UAV air combat decision making via hierarchical reinforcement learning. *Sci Rep* **14**, 4458 (2024). https://doi.org/10.1038/s41598-024-54938-5

Wray, R., Bridgman, R., Haley, J., Hamel, L., & Woods, A. (2021). Event-based keyframing: Transforming observation data into compact and meaningful form. International Conference on Artificial Intelligence, 211-221, ICAI21

Zeng, W., Zhibin, Q., Zhao, Z., Xie, C., & Lu, X. (2020). A Deep Learning Approach for Aircraft Trajectory Prediction in Terminal Airspace. IEEE Access. DOI 10.1109/ACCESS.2020.3016289