This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2020-9381C

Artificial Intelligence and Autonomy in Space:
Balancing Risks and Benefits for Deterrence and Escalation Control

Nancy K. Hayden, Kelsey Abel, Marie Arrieta, Mallory Stewart
Sandia National Laboratories

September 16, 2020

Abstract

An overarching principle accepted by space-faring nations and industry alike is to maintain freedom of operations in a safe and secure environment, commensurate with national and commercial interests. Deterrence concepts and escalation control play key roles in realizing this principle in the increasingly congested, competitive and contested space environment. AI and autonomous machine learning are being pursued as critical enablers in commercial and military programs for space traffic management, routine space operations, space domain awareness (SDA), and space control. AI systems hold the potential to strengthen deterrence by improving both the speed and ability to assess threats and inform decision makers in times of crisis. However, issues that have arisen in terrestrial AI applications will be also present in these applications, with implications for space deterrence and escalation scenarios. Key among these are performance, explainability, and vulnerability. To date there are few if any international standards or regulations to guide best practices for choosing AI methods for space operations and developing a shared understanding of the risks and benefits to strategic stability. This paper explores trade-offs between explainability, performance, and vulnerability in AI methods applied to space control and SDA scenarios, and illustrates how choices on these trade-offs may affect deterrence signaling and escalation control in space.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Introduction and Key Research Question

Freedom of operations in a safe and secure environment in space, commensurate with national and commercial interests, is a fundamental principle of international policy as well as most space-faring nations. Preventing damage to space assets is tantamount to achieving a safe and secure space environment (Defense Intelligence Agency, 2019). However, to date, there is less ability to respond to threats in space than to conventional threats, and countries with developed space assets are perceived to rely heavily on these assets for everything from civil uses to military use. As a result, deterrence concepts and escalation control play key roles in ensuring unimpeded use of space in the increasingly congested, competitive and contested space environment.

Artificial Intelligence (AI) and autonomous machine learning are being pursued as critical enablers in commercial and military programs for space traffic management (STM), routine space operations, space domain awareness (SDA), and space control (Chien and Morris, 2014; Girimonte and Izzo, 2007). AI is generally defined as methods capable of rational and autonomous reasoning, action or decision making, and/or adaptation to complex environment, and to previously unseen circumstances (Hamon and Sanchez, 2020). AI systems hold the potential to strengthen deterrence by improving both the speed and ability to assess threats and inform decision makers in times of crisis. However, issues that have arisen in terrestrial AI applications will be also present in these applications, with implications for space deterrence and escalation scenarios. Key among these are *performance*, *explainability*, and *vulnerability* – all of which can vary depending on the AI algorithms, training data, and platforms. For example, deterrence signaling might be misconstrued and responses deemed escalatory if one nation does not fully understand the intentions and strategic goals of the other nation. AI systems could negatively affect signaling by compressing the timescale for making and communicating decisions, or incorrectly classifying observed behaviors. A result could be unintentional conflict escalation.

To date there are few if any international standards and/or regulations to guide best practices for choosing AI methods for space operations and developing a shared understanding of the risks and benefits to strategic stability. This paper explores how AI deployed on critical space systems – and the design choices made about the characteristics of the AI methods - may impact deterrence signaling and escalation control. **Our key research question is, how might trade-offs between explainability, performance, and vulnerability in AI methods applied to space control and SDA scenarios affect deterrence signaling and escalation control in space?** The purpose is to stimulate dialogue on best practices for choosing AI methods for space operations and developing a shared understanding of the risks and benefits to strategic stability.

AI in Space Operations: Exemplars and Use Cases

Current and potential applications of AI in space operations are many. AI will be essential for managing mega-constellations (tens of thousands) of commercial telecommunications satellites in low Earth orbit (LEO); guiding functions such as scheduling and tasking; collision avoidance; and space debris mitigation. AI is also being explored for classification of observations from LEO constellations proposed to serve national security applications such as persistent overhead coverage and missile defense. Advancements in AI, in combination with increased availability of low-cost and secure cloud storage, have simultaneously led to improvements in SDA capabilities while decreasing costs. As databases grow with an increased number of objects to track and characterize, companies and countries will employ AI to make timely, cost-effective assessments for SDA, while reducing the role of the human-in-the-loop.

For this research, we examine two exemplars of AI applications in space, with three different use cases for each (Table 1). The first exemplar involves scenarios for threat detection and response,

Table 1 Exemplars of AI in Space and Use Cases

	Use Case A: **Immediate Threat**	**Use Case B:** **Protracted Threat**	**Use Case C:** **Crisis Breakout**
Exemplar 1: Threat Detection, Assessment, and Response			
Scenario	Space asset of Country B senses an imminent and unexpected conjunction with space asset of Country A of unknown capabilities.	Satellite from Country A tailing national security satellite of Country B in orbit; no immediate threat exists, but intentions questionable	Satellite from Country A trailing satellite from Country B begins closing orbital gap and approaching asset
AI Function	Integrate and analyze sensor input to characterize threat *(identification and classification)* and assess time available to act *(prediction);* Assess options and generate list of responses *(operational and strategic planning)* Direct action/implement command *(autonomous navigation/manipulation)*	Integrate and analyze sensor input to monitor gap between satellites *(operational planning);* Assess capabilities of Country B satellite and potential threat (*identification and classification);* Assess and recommend options to counter actions of Country B satellite without escalation *(operational and strategic planning)*	Integrate and analyze sensor input to assess threat *(identification and classification)* and time available to act (prediction); Assess and generate list of possible responses *(operational planning);* assess escalation potential of counter actions *(strategic planning);* Direct action/implement commands *(autonomous navigation/manipulation)*
Modeling Types	DES, ABM, SD	GT, ABM, SD	DES, GT, SD, ABM
Exemplar 2: Resiliency of Mega-Constellation in LEO			
Scenario	Country B experiences sudden loss of functionality of entire constellation.	A Mega-Constellation of Country B experiences cascading losses of nodes over time in a single constellation.	Country A uses anti-satellite (ASAT) capabilities against a node in the constellation of Country B, causing massive debris which threatens to interfere further with other nodes in the constellation.
AI Function	Conduct root cause assessment (*attribution*); Assess damage; Assess options and generate list of responses (to reconstitution AND potentially to retaliate)	Root cause assessment (*attribution*); damage assessment; Assess options and generate list of responses (reconfiguration, reconstitution AND signaling)	Attribute actions, predict debris field over time; conduct and enact reconfiguration accounting for future debris given risk/functionality trade-off; assess and generate suggestions for retaliatory responses (could be cross-domain)
Modeling Types	DES, ABM, SD	GT, ABM, SD	DES, GT, SD, ABM

DES—Discrete Event Simulation; SD—System Dynamics; ABM—Agent Based Model; GT – Game Theory in three different use cases that correspond to different phases of deterrence and escalation. In the scenarios for Exemplar 1, a national security satellite faces a potential threat that has yet to be realized. The primary functions of AI under these scenarios are to characterize and assess the threat, recommend (and potentially direct) options for defensive measures, and communicate a credible deterrence signal. The second exemplar involves scenarios to ensure resiliency of a meg-constellation. In these scenarios, a mega-constellation of satellites experiences failures of varying degrees. Here, primary functions of AI are to assess damage, attribute the cause, and recommend options for reconstitution and potential retaliation. These exemplars and scenarios are summarized in Table 1, along with functional roles of AI. The scenarios are viewed from the perspective of Country B.

Different types of models that might be used to explore how AI may impact deterrence dynamics are also listed in Table 1. Discrete event simulation (DES) models the operation of a system as a sequence of events over time, where each event results in a change of state in the system. DES can be helpful in setting requirements for AI in different functional roles in the use cases. Agent Based Modeling (ABM) models the system as a collection of autonomous, goal-seeking, decision-making entities. ABM can be useful for simulating the decision-making process, whether it be AI or a human-in-the-loop and interactions between decision-makers. Game theory, which can be embedded within decision-making agents, provides strategic pay-off frameworks to inform decision-making. System dynamic (SD) models represent complex systems as a network of nonlinear accumulation and feedback processes, the structure of which is designed to achieve certain goals, and whose behavior is determined by differential equations and constraints. SD simulates feedback between decisions and actions, making it particularly useful for analysis of different policy and design options for AI within each exemplar. For this initial, conceptual phase of research, we adopted an SD approach to explore system level impacts of AI design choices. We expect to incorporate additional methods (e.g., ABM and game theory) in follow-on.

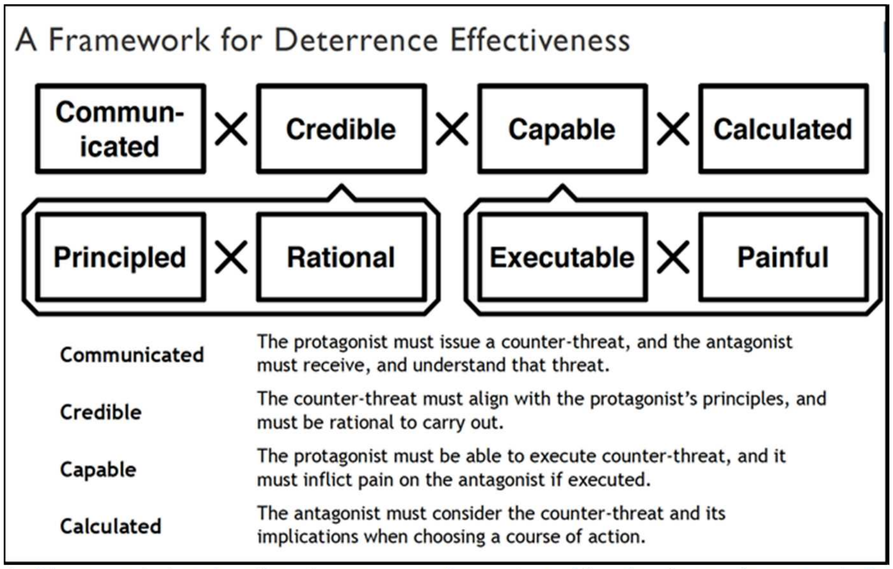
The Use Cases in Table 1 present scenarios that correspond to different deterrence phases. The first column, the Immediate Threat Use Case, contains scenarios in which a triggering event occurs without forewarning that impedes the functionality of an asset or system. This could be the result of a natural event, an accidental interaction with another country's satellite, or an unexpected attack from an aggressor. Such events require damage assessment and immediate action to resolve or mitigate damage but may not be escalatory in and of themselves. The second column, Protracted Threat Use Case, presents scenarios in which a potential threat may be perceived in the near term, requiring inference analysis, ongoing monitoring, and potential deterrence measures such as denial or dissuasion. The third column, Crisis Breakout Use Case, presents escalatory scenarios, in which a potential threat that has been monitored is being actively carried out or has just occurred. Examples could include the clearly hostile approach of a space asset or the use of Anti-Satellite (ASAT) capabilities. Such events require immediate crisis decision support and attribution, and potentially retaliation.

Table 1 also lists the functional role for AI in each of these scenarios. These roles are: identification, classification, prediction, generating responses for operational and strategic planning, and potentially the autonomous direction of action. Understanding the role of AI in each of these use cases, how humans interact with the AI, and ultimate responsibility for decisions to act to support deterrence, allows us to explore how AI systems' designs may best mitigate escalation and improve deterrence stability in space. The degree to which humans interact with the AI for decision-making, especially for retaliatory responses, will be critical in real scenarios. However, varying the degree of human-in-the-loop is outside the scope of this paper. Instead, we focus on the impact of explainability, which is a critical element for all levels of AI-Human interaction in deterrence scenarios.

Theoretical Framework: Deterrence and escalation models

We define deterrence as the ability to discourage an aggressor from taking action by either decreasing the aggressor's likelihood of success, decreasing the possible benefit the aggressor would obtain by acting, or by demonstrating that retaliation will occur if the aggressor acts. Deterrence frameworks and modeling have been discussed since the nuclear age, and there have been many different ways of conceptualizing deterrence concepts. Key to our research is the literature on models of arms races, power struggles, and deterrence (within the same domain and across domains) and escalation (including the outbreak of conflict).

We adopt the theoretical deterrence framework of Bonin and Reinhardt (2019). In this framework, effectiveness of deterrence is the product of clarity of *communication*, correctness of attackers' *calculated* risks, the *credibility* of signaled threats, and the possession of *capable* means of retaliation (Figure 1). These factors are dynamic and their interdependencies are important to understanding how events in space and responses to those events may evolve. We hypothesize that AI deployed on space systems may play a significant role in each one of these factors, based on their functional roles as described above.



Communicated — The protagonist must issue a counter-threat, and the antagonist must receive, and understand that threat.

Credible — The counter-threat must align with the protagonist's principles, and must be rational to carry out.

Capable — The protagonist must be able to execute counter-threat, and it must inflict pain on the antagonist if executed.

Calculated — The antagonist must consider the counter-threat and its implications when choosing a course of action.

Figure 1: Theoretical framework that describes deterrence as the product of four key factors (From Bonin & Reinhardt, 2019).

We draw on a number of additional modeling constructs in the literature on deterrence and escalation to instantiate this framework into a conceptual model. Table 2 shows a summary of this literature, and the relevant deterrence and modeling concepts; These are a select subset of the literature that we deem most relevant for the understanding and formulation of our conceptual model of potential roles of AI in deterrence and escalation dynamics.

Table 2 summary of key literature that seeks to model deterrence and escalation concepts.

Model	Methods	Features of Deterrence
Richardson Arms Race (1960)	DE	Seeks to understand arms races, how they arise and perpetuate, and their impact on the likelihood that conflict will occur. One nation's investment in capabilities is dependent on the arms of the adversary, the fatigue of the nation itself, and the history between the two nations.
Mauro's Mixed-Strategy Models of Conflict (2016)	SD, GT, LL	This model predicts that signaling, perception of the adversary, and dominant strategies greatly influence the outcome of a conflict. It also creates a valuation of payoffs using dollar amounts and expected loss per engagement which could be used by an AI to determine the risk/benefit payoff of a given action or series of actions during a potential threat or disruption.
Lutijen's Major Power Escalation (n.d.)	ABM, GT	Conflicts become more likely when power parity exists, when satisfaction with the status quo is low, and when one's power is not in decline
Lanchester Model of Fighting War, Lanchaster Square Law (1916)	DE	If one force outnumbers its opposition, its effective firepower is the square of the total number of units in the larger force. As a result, the smaller force should engage the larger in limited arenas and focus their attack into a limited part of the larger force.
Intriligator and Brint (1984)	GT, DE	Stable deterrence exists when costs of conflict are sufficiently high, and dynamic changes occur when capabilities suddenly expand.

DE–Differential Equations; GT–Game Theory; SD–System Dynamics; LL–Lanchester Laws; ABM–Agent Based Modeling

The Lanchester Laws (1916) provides a lens through which to look at a nation's *credibility* and an antagonist's *calculation* and helps explain why threats in space are increasing. Utilizing differential equations, these laws predict that if one force outnumbers its opposition, its effective firepower is the square of the total number of units in the larger force. As a result, the smaller force should engage the larger in limited arenas and focus their attack into a limited but critical part of the larger force. This conclusion is relevant within the space domain as powerful countries are perceived to rely heavily on their space assets (Defense Intelligence Agency, 2019), yet have less ability to respond to threats in space than to conventional threats. Thus, the Lanchester Laws predict that nations or entities with less power may *calculate* that they should act against more powerful countries in the space domain where aggressors can focus their efforts on high value targets that are harder to defend. This incentive is augmented by the decreasing barriers to enter, navigate, and act in space, further enabling aggressors to target space assets. By bolstering a countries' ability to detect, assess, and respond to threats, usage of reliable AI systems would make evasive responses in space more *credible* and would decrease the likelihood of success for would-be aggressors, thereby deterring threats on space assets. The more *painful* a potential reposnse might be, the more important it will be to ensure accountability for *principled* reponses, which may require an accountable human-in-the-loop. This dynamic behavior between AI *capabilities*, *credibility* of crisis management, and *calculation* of aggressor's actions is pivotal to modeling how a nation would act and interact in the space domain and can be further enhanced with concepts like game theory.

Intriligator and Brint (1987) did just this, combining differential equations and game theory to understand when and how deterrence holds. Each actor has some threshold under which they would not initiate conflict because the benefits are too small, and another threshold over which they would not initiate conflict because the potential costs are too large. Figure 2 shows these thresholds and the resulting regions of the graph where deterrence holds, where deterrence is uncertain or unstable, and where one actor can successfully make the other comply. Intiligator and Brint also show how the region of the graph the actors find themselves in changes dynamically as *capabilities* change. The arrow labeled (2) in Figure 2 demonstrates the pathway that would be traversed if entity j installs a new capability that entity i cannot effectively respond to and deter, thereby shifting the balance of power in j's favor and increasing the odds of conflict. We postulate that this pathway can lead to a Richardson-esque arms race, and that the two models can be combined to understand dynamic behavior in the system.

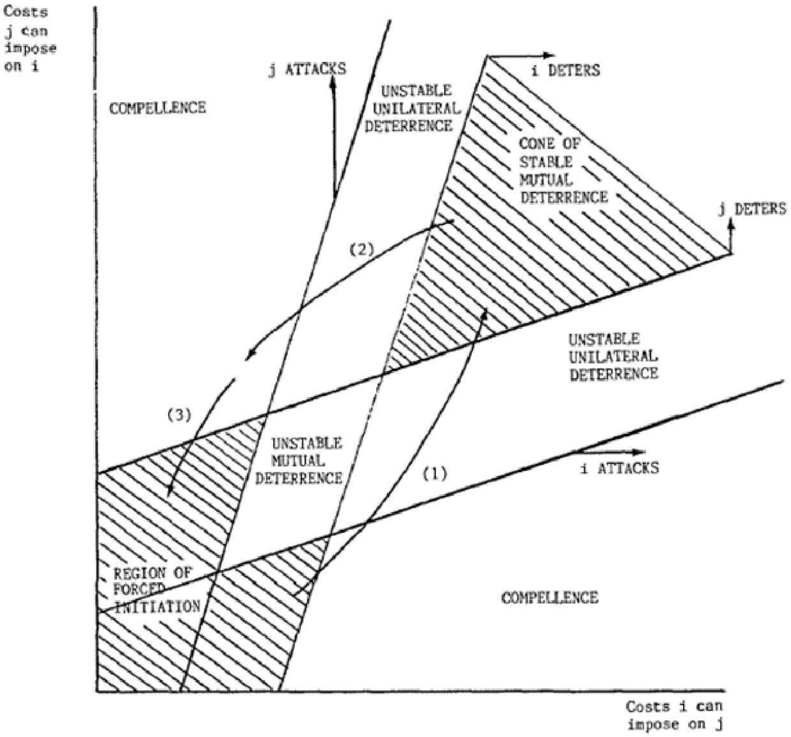

Costs j can impose on i
COMPELLENCE
j ATTACKS
UNSTABLE UNILATERAL DETERRENCE
i DETERS
CONE OF STABLE MUTUAL DETERRENCE
j DETERS
(2)
UNSTABLE UNILATERAL DETERRENCE
(3)
UNSTABLE MUTUAL DETERRENCE
(1)
i ATTACKS
REGION OF FORCED INITIATION
COMPELLENCE
Costs i can impose on j

Figure 2: A graphical view of deterrence demonstrating regions of compellence, stable deterrence, and unstable deterrence (From Intrilligator and Brito, 1987).

Hypotheses for AI impacts on deterrence

AI methods are constantly evolving and improving with new and complex algorithms being routinely developed by researchers for competitive advancements across many different fields within public and private sectors, including the space domain. Table 3 summarizes some common types of algorithms and methods used for AI, the primary function of each method and its utility, and data needs.

Table 3 AI Algorithm Types, Functions and Data Needs

Algorithm Type	Utility	Data Needs
Supervised classification based on pre-determined categories		
Support Vector Machine (Vapnik, 1995)	PREDICT a category/IDENTIFY trends, sentiments, fraud, or threats	Sensor data, numeric data/ streaming text data, static text data
Discriminant analysis (McLachlan, 2004)	PREDICT an output based on historical and current data	Continuous sensor, numeric data
Naïve Bayes (Russell and Norvig, 2020)	PREDICT a category, IDENTIFY trends, sentiments, fraud, or threats	Sensor data, numeric data/ streaming text data, static text data
Nearest Neighbor (Russell and Norvig, 2020)	PREDICT an output based on historical and current data	Sensor data, numeric data/ streaming text data, static text data
Supervised regression based on correlation		
Linear regression, Generalized Linear Models (Russell and Norvig, 2020)	IDENTIFY trends, sentiments, fraud, or threats	Sensor data, numeric data/ streaming text data, static text data
Support Vector Regression, Gaussian Process Regression	PREDICT a quantity	Sensor data, numeric data/ streaming text data, static text data
Ensemble methods	PREDICT a quantity	Sensor data, numeric data
Decision trees	PREDICT an output based on historical and current data	Sensor data, numeric data
Neural networks	MOVE an object physically or in a simulation	Sensor data, numeric data, Mathematical models, videos, lidar
Unsupervised clustering based on observed patterns		
K-means, K-medoids Fuzzy C-Means	IDENTIFY objects or actions in image, video, and signal data, Anomaly detection	Images, videos, signals
Hierarchical Clustering	IDENTIFY objects or actions in image, video, and signal data	Images, videos, signals
Gaussian Mixture	IDENTIFY objects or actions in image, video, and signal data	Images, videos, signals
Neural Networks	IDENTIFY objects or actions in image, video, and signal data and ENHANCE images and signals and RESPOND to speech or text commands based on context and learned routines	Images, videos, signals
Hidden Markov Model	Compute probability of sequence of observable events (PREDICT), including those we infer because cannot observe directly.	Images, videos, signals
Reinforcement Learning based on goal-oriented interactions with environment		
Deep neural networks (DNNs)	Determines optimal behaviors based on past experiences and current state of the environment (PLANNING), can model complex, non-linear relationships without a significant amount of marked data.	Large data sets on actions and outcomes.

The AI methods (algorithms types) in Table 3 are grouped into four categories: supervised classification, supervised regression, unsupervised clustering and reinforcement learning. Supervised algorithms require labeled training data to help predict outcomes for unforeseen data. Unsupervised algorithms can find unknown patterns and features in unlabeled data useful for categorization, and are used for more complex problems, (Filiz, 2017). Finally, reinforcement learning describes a class of problems where an agent operates in an environment without an initial training dataset. The agent learns to map situations to actions and the feedback loop is between the agent's actions and the environment, (Sutton and Barto, 2018). Some of the recent advancements in reinforcement learning employ combinations of efficient learning algorithms that explore huge parameter spaces (e.g., hundreds of layers and millions of parameters) in complex "black-box" DNN models (Arrieta et al, 2019).

An AI system may involve many of these methods, each with a different function relative to space control and/or SDA. Issues that have arisen in terrestrial AI applications will be also present in these applications, with implications for space deterrence and escalation scenarios. Key issues for deterrence and escalation are *performance*, *explainability*, and *vulnerability*. Specific characteristics depend on the context and circumstances of each application, and whether those conditions are changing over time. As a general rule of thumb, all else being equal, current reinforcement learning methods can be expected to have the highest performance but lowest explainability; while supervised classification methods will have higher explainability but lower performance, and higher vulnerability. Current unsupervised learning methods are typically somewhat in between the two (Figure 4). These general assumptions may change, however, if a situation evolves outside of the intial training set.

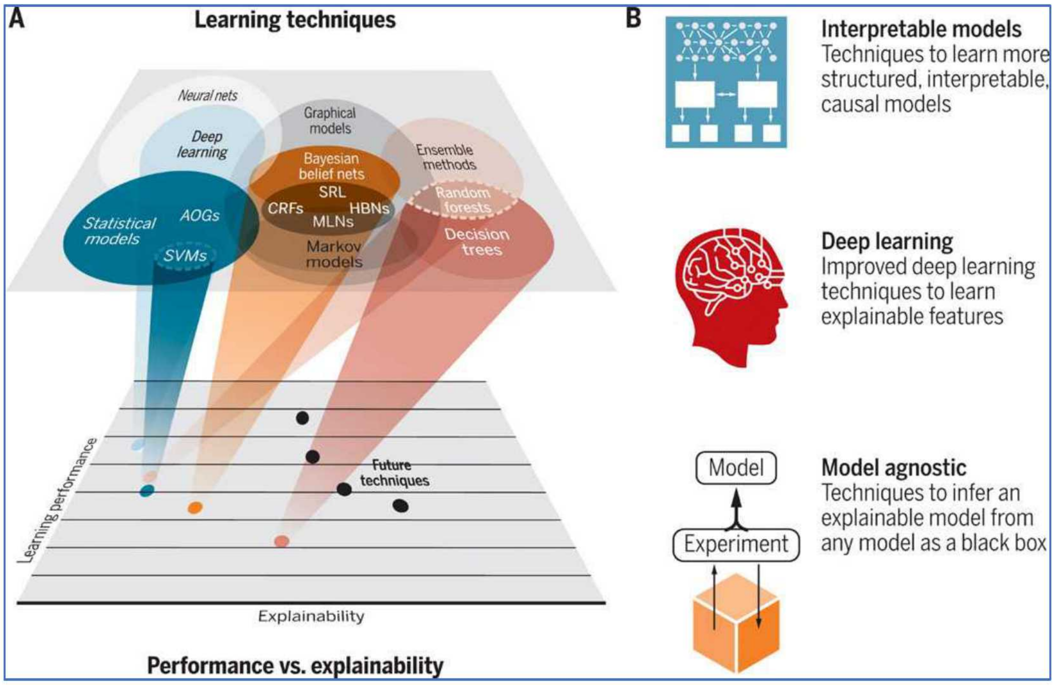


Figure 3 Trade-offs between performance and explainability (from Gunning et al 2019). Reprinted with permission.

We hypothesize that these issues will affect deterrence effectiveness, as conceptualized in Figure 1, in the following ways:

Hypothesis 1: ***All else being equal, AI systems with higher performance metrics for correctly identifying, classifying and responding to a threat in space in a timely manner, will increase deterrence credibility.***

The performance of AI methods (e.g., accuracy, precision, recall, sensitivity, latency, confusion, rate of learning) is evaluated by a user-defined procedure and a set of metrics that ultimately should assess the AI's ability to solve a posed problem (Sunasra, 2017). Ultimately, the credibility of an AI’s response to a given attack is dependent on how well the AI performs a given task, (identification, categorization, planning).

Hypothesis 2: ***All else being equal, AI systems with higher explainability will reinforce deterrence credibility and communications for deterrence signaling. AI systems with lower explainability increase risk of escalation, and loss of external perceptions of legitimacy.***

AI explainability, as defined by Montavon (2018), is the ability to present a collection of features that an AI algorithm has used to produce output, in terms understandable to humans. Moreover, we adopt the convention that explainability is an active characteristic of a model (e.g., action or procedure), distinct from the passive characteristics of interpretability or transparency, and depends on the audience (Arrieta, 2019). For example, explainability provides a decision-maker with information necessary to understand why the AI has classified certain situations as hostile or neutral, and why the AI recommends a specific action (or series of actions) in a particular context. This enables the decision maker to understand, and potentially trust, an AI system overall, and to communicate rationale for the decisions to others, thereby enabling transparency. Without explainability and transparency, there may be misinterpretation of signals, and increased risk of inadvertent or disproportional engagement.

Hypothesis 3: ***Higher AI vulnerability will decrease deterrence credibility and confidence***

Vulnerabilities in AI may lead to malfunctions that occur naturally in the course of program execution, or intentionally introduced by an adversary in an algorithm (or model) that otherwise performs well (Hamon et al, 2020). Typical vulnerabilities of AI systems include data poisoning, the crafting of patterns for classifying adversarial behaviors, and exploitation of known weaknesses in the learning process (e.g., sensitivity to noise).

Contrary to traditional cyberattacks that exploit gaps or mistakes in the underlying code, AI attacks are enabled by the limitations to the underlying machine learning techniques powering the system. AI systems utilize these algorithms to extract information and generalize patterns from data; the patterns are stored within the model and continuously updated as new datasets are input. Over time, robust patterns are learned that allow the AI to outperform humans on many tasks; however, the complete dependence of AI methods on the dataset exposes the vulnerability of AI methods to

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525. SAND2020-XX

attack or tampering. If the dataset is the AI model's only source of knowledge, the data itself can be attacked through the purposeful introduction of aberrations in the data.

Adversarial attacks and/or deception through these means may result in misperception of a situation as hostile, misattribution, and/or miscalculation of appropriate response.

Hypothesis 4: ***Stable equilibrium (e.g., balance between deterrence and escalation) depends on the trade-off between performance, vulnerability, and explainability.***

Table 4 Hypotheses for Effects of Performance, Vulnerability, and Explainability Trade-offs on Deterrence and Escalation

	Performance	**Vulnerability**	**Explainability**	**Deterrence Impact**	**Escalation Impact**
i	High	High	Low	Decrease	Increase
ii	Low	High	High	Decrease	Decrease
iii	High	Low	Low	Increase	Increase
iv	High	Low	High	Increase	Decrease

Due to the inherent tradeoffs between *performance (P)*, v*ulnerability (V)*, and *explainability (E)*, it is not possible for optimize all three at once. Moreover, different traits support different goals; high explainability may serve to build trust of the decision makers who rely on them, but more explainable models may be less capable of handling complexity and therefore performance may suffer (Gunning, 2019). A high performing model (e.g., one that rapidly incorporates new data into current models), may also have high vulnerability to spoofing. Making decisions based on patterns found in anomalies in the input data could lead to high consequence actions, potentially increasing escalation (Hamon, 2020). Conversely, a high performing model that has lower vulnerability could adversely affect escalation if the inputs and decisions made by the AI are not explainable and transparent. In only one case do the combination of characteristics both increase deterrence (e.g., reduce threats) while decreasing escalation. However, this combination of AI characteristics is not technically feasible at this point in time.

The inherent trade-offs of these characteristics create challenges for deterrence, which include building confidence in decision-making, and effective communications for signaling. Confidence can increase the speed at which decisions can be made. However, war games have demonstrated that the speed of AI systems can lead to inadvertent escalation, due to incorrectly classifying observed events and interpreting signals, leading to disproportionate response (Wong, 2020). Table 4 summarizes our qualitative estimations for how combinations of PV&E affect deterrence and escalation.

Analytic Approach: System Dynamics Model

The dynamic deterrence concepts, modeling approaches and AI characteristics described in the previous sections are combined into the system dynamics model in Figure 5. The model is a Causal Loop Diagram (CLD) for exploring the research question, ***how might trade-offs between explainability, performance, and vulnerability in AI methods applied to space control and SDA scenarios affect deterrence signaling and escalation control in space*?** This CLD was constructed to be representative of dynamics in both exemplars in Table 1, from the perspective of Country B, and demonstrate and test the four hypotheses above.

CLD Model Structure and Dynamics

There are two dominant feedback loops in the CLD model: deterrence and escalation. The relative strength of these loops depends on the levels of the four primary stocks (e.g., variables that accumulate over time): DETERRENCE CREDIBLITY, HOSTILE THREAT PERCEPTIONS, SPACE CONTROL CAPABILITIES, and EXTERNAL PERCEPTIONS OF LEGITIMACY. The levels of these stocks are determined, in turn, by feedback between the dynamic variables, **hostile threat rate, and effectiveness of communications**, where communications may include both signaling and explanations. The system is in stable equilibrium when DETERRENCE CREDIBIILTY and EXTERNAL PERCEPTIONS OF LEGITIMACY work together to balance **hostile threat rate.**

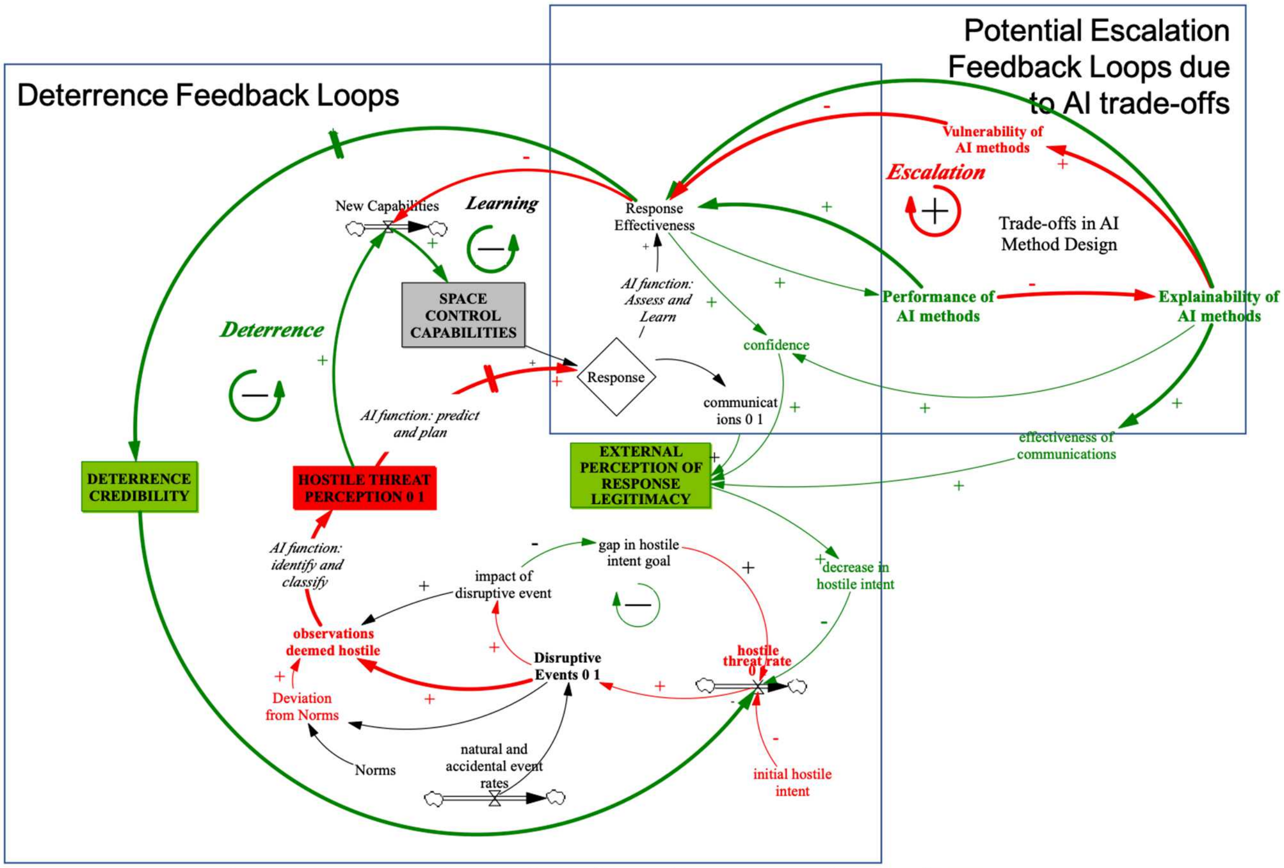

Potential Escalation Feedback Loops due to AI trade-offs
Deterrence Feedback Loops
Vulnerability of AI methods
Escalation
Trade-offs in AI Method Design
New Capabilities
Learning
Response Effectiveness
AI function: Assess and Learn
SPACE CONTROL CAPABILITIES
Performance of AI methods
Explainability of AI methods
Deterrence
confidence
Response
communicat ions 0 1
AI function: predict and plan
effectiveness of communications
DETERRENCE CREDIBILITY
HOSTILE THREAT PERCEPTION 0 1
EXTERNAL PERCEPTION OF RESPONSE LEGITIMACY
AI function: identify and classify
gap in hostile intent goal
impact of disruptive event
decrease in hostile intent
observations deemed hostile
Disruptive Events 0 1
hostile threat rate 0
Deviation from Norms
Norms
natural and accidental event rates
initial hostile intent
+
-

Figure 4: Causal Loop Diagram (CLD) of AI used in space operations and deterrence.

SPACE CONTROL CAPABILITIES is a stock with an initial value that increases proportionally to the rate of **new capabilities**. The rate of increase in **new capabilities** is proportional to the level of HOSTILE THREAT PERCEPTIONS, and inversely proportional to **Response Effectiveness**. In the exemplars for this paper, SPACE CONTROL CAPABILITES affects **Response Effectiveness** based on what capabilities exist on or near the targeted satellite (Exemplar 1) or the

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2020-XX

resilience of the constellation (Exemplar 2). A secondary feedback loop of **Learning** from **Response effectiveness** creates **new capabilities**. As **new capabilities** increase, the benefit an actor could receive by attacking is diminished (e.g., deterrence by denial), thereby decreasing the number of hostile actions that are "worth" attempting.

Effectiveness of Communications—between defender (Country B) and aggressor (Country A), between the AI and decision makers, and between Country B and the international community—increases the EXTERNAL PERCEPTIONS OF RESPONSE LEGITIMACY and improves Country B's ability to deter and respond to events. **Confidence** is a function of **response effectiveness** and **explainability of AI methods**. **Confidence** increases EXTERNAL PERCEPTIONS OF RESPONSE LEGITIMACY and decreases the **hostile threat rate** by decreasing the potential payoff to would-be aggressors. DETRRENCE CREDIBILTY is an endogenous stock which is dependent on **response effectiveness** and **explainability of AI methods,** and which influences the rate of **hostile threats**. **Response effectiveness** depends on the **response** type chosen (e.g., denial or punishment), and the SPACE CAPABILITIES for carrying it out. Each time the **Response Effectiveness** is high, then the defender (Country B) adds to a track record of being well-equipped to repel attacks and be resilient, thereby increasing the stock of DETERRENCE CREDIBLITY. This, in turn, reduces **hostile intent** by influencing the calculus of the aggressor (Country A) about whether or not to engage in an attack on a space asset or system.

The Intriligator and Brito deterrence landscape provides insights into the dynamic behaviors that could emerge from this model. In equilibrium (e.g., when **hostile threat rate** is low) two adversarial nations are in either the cone of mutual deterrence (a situation in which both nations have sufficient capabilities and resiliency as to mutually deter action against space assets) or the Region of Forced Initiation (a situation in which both nations lack hostile intent to make a challenge in space or both lack capabilities to cause disruption to space systems). However, when country A initiates a hostile action in space (e.g., **disruptive event** the model), both countries enter the unilateral unstable deterrence region. From there, Country B can either choose if, when, and how to respond.

Whether or not Country B responds will be determined in part by HOSTILE THREAT PERCEPTION and SPACE CONTROL CAPABILITIES. No response would put the two countries into the Compellence region of Figure 5, where Country A would receive a benefit and elevated status while country B would lose deterrence credibility. If country B chooses to respond, then the region the countries would find themselves in would be dependent on the proportionality of B's actions. A weak response from Country B (low **operational responses effectiveness** in Figure2) would shift the countries into the unstable mutual deterrence region; an overly harsh response (low EXTERNAL PERCEPTION OF LEGITIMACY) would shift the countries into the region of unstable unilateral deterrence favoring country B, which would be considered an escalation and may cause Country A to increase **hostile threat rate**; a proportional response from Country B (high EXTERNAL PERCEPTION OF LEGITIMACY and HIGH DETERRENCE CREDIBILITY) would shift the two countries into the cone of mutual deterrence and increase the likelihood that deterrence will hold in the future.

Additional dynamics within the CLD model that might reduce deterrence are:

- **Operational response effectiveness** is low, leading hostile actors to conclude that they can “get away” with more and therefore **hostile threat rate** increases.
- Limited SPACE CONTROL CAPABILITES constrain **operational response effectiveness** and **new capabilities** rates are low, weakening **deterrence** feedback loop and resulting in an increase in **hostile threat rate**.
- **Effectiveness of communications** is low, reducing EXTERNAL PERCEPTION OF LEGITIMACY and possibly leading to an increase in **hostile threat rate.**

The dynamic processes represented in the CLD model in Figure 5 are analogous to the OODA Loop decision making model. The OODA Loop - Observe, Orient, Decide, Act - is a tool that enables rapid, adaptable decision making. It is used by military strategists to conduct combat operations by enabling decision makers to take limited information, make the best decision possible (operational or strategic), and adjust the plan when further issues arise (Osinga, 2005). When a disruptive event is observed, AI enables three critical functions for the OODA loop: (1) supports Observing and Orienting by identifying and classifying **disruptive events** (e.g., as hostile, natural or accidental); (2) supports Deciding and Acting by predicting short and long-term effects of the event, generating options for **response**, and recommending actions to meet pre-determined goals; and (3) supports OODA loop speed by assessing the **response effectiveness** and learning how to improve performance for future situations, both operationally and strategically.

Hypothesis Testing and Implications

As an initial test of our hypotheses for how different *PV&E* characteristics of AI systems lead to different outcomes of deterrence and escalation, we demonstrate how the dynamics of one of the Use Cases in Table 1 play out in our model (Figure 5), assuming different values for PV&E. We examine Use Case B of Exemplar 1, in which a satellite from Country A is tailing a national security satellite of Country B in orbit. No immediate threat exists, but the intentions of Country A are uncertain. Each of the following scenarios applies a different set attributes of an AI system and traces how the outcomes change based on the design of the AI system.

Reference Scenario (High P, Low V, High E): The reference scenario demonstrates Hypothesis (iv) in Table 4. Using supervised learning classification algorithms (e.g., Support Vector Machine, discriminant analysis), the AI system on Country B’s satellite accurately observes the **disruptive event** of the trailing satellite and its **deviation from the norm** with a high degree of **confidence**. Using a combination of classification and regression techniques, the system **classifies** the identity of the asset as Country A with high confidence. There is currently no immediate threat, but should an attack occur, the **impact of this disruptive event** would be high. Using supervised neural networks, the AI predicts that actions must be taken, and provides a recommended plan, including response and signaling options. The decision maker decides on a **response** of slightly altering the trajectory of the space asset to determine if the trailing is accidental or intentional. The AI then enacts this change and continues to monitor path data. Country A’s asset continues on course, and the two satellites do not interact further. Having correctly identified the **disruptive event** early, the AI allowed decision makers to improve their **Response Effectiveness** and remove their asset from close proximity to the trailing satellite that could have potentially approached and harmed the space asset. Country A, having received a signal and having witnessed the AI assess the problem and take corrective actions, now perceives Country B's response effectiveness as more robust, which increases Country B's EXTERNAL PERCEPTION OF LEGITIMACY, and DETERRENCE CREDIBILITY.

Outcome: Deterrence and escalation decrease.

Scenario A (High P, High V, Low E): This scenario demonstrates Hypothesis (i) in Table 4. Using unsupervised learning classification algorithms, the AI system on Country B's satellite observes the tailing satellite and predicts that the two satellites are on ultimately diverging orbits. The data on which this prediction is made has been intentionally manipulated by Country A to "spoof" classification and prediction algorithms based on known weaknesses in the learning process (e.g., sensitivity to noise). Commercial space operators detect and publicly report on the object trailing the national security satellite, speculating on **hostile intent**. The AI system is updated with corrected information, attributes **hostile intent** to Country A, and predicts what changes in orbit would separate the two satellites while maintaining mission operations. AI enacts a defensive **response** - altering the trajectory of Country B space asset, while planning ahead for how that might be received, as this new trajectory will bring Country B's asset into close proximity with an asset of Country A. It is unclear whether the intention of County B is to pose a threat to the new asset of Country A, as the explainability of the AI system is low. Country A ceases to tail Country B asset, but raises the alert status for perceived threat level from Country B

Outcome: Country B DETERRENCE CREDIBILTY is reduced. Country A **hostile intent** and escalation increases due to lack of **effective communication** regarding behind Country B's maneuvers.

Scenario B (Low P, High V, High E): This scenario demonstrates Hypothesis (ii) in Table 4. Country A interferes with sensors aboard Country B satellite, resulting in a skewed data set for the AI identification and classification systems of Country B. This is compounded by introduction of sources of poisoned data regarding operations of Country A satellite. Country B's satellite therefore does not classify presence as a **hostile threat**. Commercial space operators detect and publicly report on the object trailing the national security satellite, speculating on **hostile intent**, but without attribution to Country A. The AI system is updated with corrected information and predicts what changes in orbit would separate the two satellites while maintaining mission operations. AI enacts a defensive **response** - altering the trajectory of the space asset, while planning ahead for how that might be received, using supervised classification and regression techniques. As this new trajectory will bring Country B's asset into close proximity with an asset of Country A, AI anticipates questions about the maneuver, and sends communications to explain the maneuver. In this case, the **Response Effectiveness** is initially low due to performance and vulnerability issues with data, improving only after supplemental data is received from commercial operators, which is reported publicly. Country B's DETERRENCE CREDIBIITY decreases as a result. However, Country B's EXTERNAL PERCEPTION OF LEGITIMACY increases due to the explainability of the eventual response.

Outcome: Information loss with delayed response effectiveness. Deterrence decreases. Escalation also decreases with explanation of defense responses.

Scenario C (High P, Low V, Low E): This scenario demonstrates Hypothesis (iii) in Table 4. Using supervised learning classification algorithms (e.g., SVM, discriminant analysis), the AI system on Country B's satellite accurately observes the **disruptive event** of the trailing satellite and its **deviation from the norm** with a high degree of **confidence**. Using a combination of classification and regression techniques, the system **classifies** the identity of the asset as Country A with high confidence. There is currently no immediate threat, but should an attack occur, the **impact of this disruptive event** would be high. Using Deep Neural Networks, the AI predicts that actions must be taken, and provides and implements a plan, including response and signaling options based on feedback from the environment. The initial **response** of Country B is to slightly alter the trajectory of the space asset and observe the response of Country A. Country A's asset changes course and continues to trail Country B asset. The AI plan directs a secondary maneuver which puts Country B asset in close proximity to another asset of Country A. It is unclear whether the intention of County B is to pose a threat to the new asset of Country A. Country A ceases to tail Country B asset, but raises the alert status for perceived threat level from Country B. In this scenario, the explainability of the AI to both County A and Country B, and the degree to which there is a human-in-the-loop, will be important for demonstrating and **communicating** adherence to principles for rules of engagement.

Outcome: Deterrence increases; escalation increases. Country B successfully deters immediate aggression from Country A, but Country A escalates threat level due to lack of explanation of intent behind Country B's maneuvers.

Insights and Future Work

This work addresses the importance and increasing usage of AI as an enabler of space systems, as well as the potential of AI to facilitate decision making during accidental, natural, or hostile crises in space. To better understand the impact of AI systems to either enhance the ability to deter threats in space or increase the likelihood of conflict escalation, a CLD model of system dynamics was constructed. This model captures hypotheses regarding AI's impact on identifying, assessing, and responding to threats in space, taking into account both the OODA Loop and the tradeoff characteristics of AI systems (*performance, explainability*, and *vulnerability*). The value proposition for AI is to improve operational and strategic planning quality in complex and ambiguous situations, while increasing speed of decisions, thereby strengthening deterrence. However, AI may also trigger unintended escalation. The outcome depends in part on the trade-off in performance, explainability, and vulnerability of AI systems. Understanding how these factors influence the overall effectiveness of operational responses—either to strengthen or impede decision making—is an important step in designing protocols for AI use in decision making when space assets are threatened. Robust protocols may require AI on orbit for years, learning from feedback given by human operators.

This preliminary analysis is limited in several ways. The CLD model represents the decision-making process of only one entity; and does not include explicit models of decision-making. Both limitations could be overcome in future work by extending this model to include multiple decision-making perspectives (and design choices for AI), represented through ABM and Game Theory. This would allow us to explore how different types of AI and different levels of AI use by different actors affect the equilibrium dynamics and ultimate deterrence and escalation outcomes.

Future analytic work could (1) create a simulation model from the CLD for quantitative analysis of AI design trade-offs; (2) expand the CLD to incorporate two actors represented by an Agent Based Model to simulate how two or more entities using different AI systems (and therefore with different values of explainability, vulnerability, and performance) might behave when interacting with one another. Included would be different approaches to explainability, contrasting, for example, methods based on information processing theory versus philosophy of explanations from the social sciences (Mittlestadt et al, 2019). Game theory could also be incorporated to understand ways that AI space systems could step through a dynamic situation when a possibly escalatory action is taken and understand what effects AI systems would have on the overall outcome of the skirmish. Such a simulation model can be run stochastically to explore variations of performance, explainability and vulnerability parameters and approaches across use cases, with the ultimate goal of developing a deterrence phase diagram similar to the conceptual model of Intrilligator and Brito.

To operationalize the utility of the CLD and our modelling concepts more broadly, we need to step back and confirm agreement in the AI community with some of our basic assumptions. We need to confirm with other AI users that there are both risks and benefits to deterrence and escalation control through the very use of AI in space assets. Beyond that, we need to confirm that performance, explainability, and vulnerability are relevant (if not the most relevant) parameters for many of the decision to incorporate AI technologies into space systems. We can accept that these are relevant parameters for AI method decision-making, then we also need to confirm that these words mean the same things to different user communities. For example, what may be deemed successful performance, or sufficient explainability (for trust-building purposes) in a certain space faring community may not be the same in others. The AI rules of engagement must be transparent to all nations to avoid escalation. Moreover, the comfort level with different amounts of vulnerability may depend on the reliance on the particular space system in which the AI is utilized. A recent National Security Commission on Artificial Intelligence (NSCAI) Report suggested different parameters "for creating and maintaining trustworthy and robust AI systems." We feel that performance, explainability, and vulnerability are a subset of the NSCAI recommended parameters, which include auditability, traceability, interpretability, and reliability. (NSCAI, 2020).

Important outcomes of this work include stimulating dialogue among space actors about how AI use may impact the stability of deterrence in space. Recognizing these are requirements in order to operationalize and even explain the CLD, we suggest greater collaboration and communication between the relevant international communities relying on AI to support their space assets. If there was general consensus on both the challenges AI presents to deterrence and escalation control and the advantages AI offers, the broader user community could discuss common methods and best practices to minimize these challenges and maximize these advantages. If the meaning of performance, explainability, and vulnerability could be universalized, then we could at least start to establish threshold levels for each of those terms that could be worked into the AI methods in use in space systems. In other words, there could be a base level of performance that all AI users would meet in order to minimize accidents; a base level of explainability that the AI would have to have in order to minimize miscommunication; and a base level of vulnerability that all AI users would understand, accept, and share for the purposes of transparency and greater engagement opportunities. All of these thresholds must be interrelated and interdependent, just as stakeholders in space are not independent. The creation of space debris by mishap or malfeasance affects all space players. If the space faring community could agree on certain requirements for operating AI systems in space, they may be able to minimize miscommunications, misperceptions, and malfeasance. Through minimizing these, AI's advantages for deterrence and escalation control could be maximized.

References

Arrieta, Alejandro, Natalia Rodriguez, and Javier Del Ser, “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”, *Information Fusion*, November 2019. DOI: 10.1016/j.inffus.2019.12.012

Ben-Hur, Asa, David Horn, Hava T. Siegelmann, and Vladimir Vapnik (December 2001), “Support Vector Clustering”, *Journal of Machine Learning Research* (2) pp 125-137. https://www.jmlr.org/papers/volume2/horn01a/horn01a.pdf

Bonin, Ben and Jason Reinhardt, “Deterrence & Nuclear Strategy: An Imperfect Overview” (2019). Sandia National Laboratories, Albuquerque, NM. SAND2019-9183 PE.

Caspary, William R. (1967, March) Richardson’s Model of Arms Races: Description, Critique, and an Alternative Model. *International Studies Quarterly*. 11(1) pp. 63-88. DOI: 10.2307/3013990, https://www.jstor.org/stable/3013990

Chien, Steve and Robert Morris (2014) Space Applications of Artificial Intelligence, *AI Magazine*, Winter 2014, Association for the Advancement of Artificial Intelligence. ISSN 0738-4602.

Defense Intelligence Agency (2019, January) Challenges to Security in Space. Washington, DC.

Fahrettin Filiz, Fahrettin, “Artificial Intelligent Algorithms,” (2017). Retrieved from: https://towardsdatascience.com/4-1-artificial-intelligent-algorithms-aff1a1ca910a

Gillespie, John V., Zinnes, Dina A., Tahim, G. S., Sampson III, Martin W., Schrodt, Philip A., and Rubison, R. Michael. (1979) Deterrence and Arms Races: An Optimal Control Systems Model. *Behavioral Science*. 24(4), 250-262.

Girimonte, D. and D. Izzo (2007). Artificial Intelligence for Space Applications. Intelligent Computing Everywhere. A. J. Schuster. London, Springer London: 235-253.

Gunning, D., Mark Stefik, Jaesik Choi, Timothy Miler, Simone Stumpf, and Guang-Zhong Yang (2019). *XAI- Explainable artificial intelligence*, Science Robotics **4**, eeay7120. 18 December 2019.

Hall, Patrick and Navdeep Gill (2018). An Introduction to Machine Learning Interpretability, O’Reily Media Inc. Sebastopol, CA.

Hamon, R., Junkelwitz, H. and Sanchez, I. (2020). *Robustness and Explainability of Artificial Intelligence: From technical to policy solutions*. EUR 30040, Publications Office of the European Union, Luxembourg, Luxembourg, 2020, ISBN 978-92-79-14660-5 (online), doi:10.2760/57493 (online), JRC119336.

Hill, Walter W. (1992) Several Sequential augmentations of Richardson’s arms race model. *Mathematical and Computer Modeling*. 16 (8-9): p 201-212. https://doi.org/10.1016/0895-7177(92)90096-4 Retrieved from https://www.jstor.org/stable/3013990?seq=1

Intriligator, Michael D. and Brito, Dagobert L. (1981) Nuclear Proliferation and the Probability of Nuclear War. *Public Choice* 37:247-260. Martinus Nihoff Publishers, The Hague, Netherlands.

Intriligator, Michael D. and Brito, Dagobert L. (1984, March) Can Arms Race Lead to the Outbreak of War? *Journal of Conflict Resolution*. 28 (1): pp. 63-84.

Ishida, Atsushi (2015) An Initial Condition Game of Richardson’s Arms Race Model. *Sociological Theory and Methods*. 30(1), pp. 37-50.

Khan, M., Usman, F., and Khichi, A. (n. d.) Richardson Arms Race Model. Powerpoint retrieved from http://www.mtholyoke.edu/~tchumley/m333/Richardson-Arms-Race.pdf

Kugler, J. and Zagare, F. C. (1989, March 24) The Long-Term Stability of Deterrence. *International Interactions*. 15(3/4): pp. 255-278. Retrieved from https://www.researchgate.net/publication/248937604_The_Long_Term_Stability_of_Deterrence

Lanchester, F. W. 1916. Aircraft in warfare: The dawn of the fourth arm. Constable limited.

Lutejin, R. F. M. (n. d.) Major Power Wars: a Model Based Exploratory Analysis.

Mittelstadt, Brent, Chris Russel and Sandra Wachter (2018), "Explaining Explanations in AI", Proceedings of FAT* 19 Conference on Fairness, Accountability and Transparency (FAT*19), January 29-31, Atlanta GA. doi/10.1145/3287560.3287574. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3278331. Posted November 2018.

Montavon, G., W. Samek, and K.-R. Muller, "Methods for interpreting and understanding deep neural networks," Digit. Signal Process., vol. 73, pp. 1-15, Feb. 2018.

Mauro, Lou (2016, April) Evolutionary Mixed-Strategy Models of Conflict. Strategy Modeling and Business Dynamics Worchester Polytechnic Institute—term project.

Mittelstadt, Brent, Chris Russell, and Sandra Wachter (2019). "Explaining Explanations in AI". In *FAT 19: Conference on Fairness, Accountability, and Transparency (FAT 19*), January 29-31, Atlanta, GA. ACM, NY, NY. https://doi.org/10.1145/3287560.3287574

McLachlan, G.J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience.

National Security Commission on Artificial Intelligence "Key Considerations for Responsible Development & Fielding of Artificial Intelligence, Abridged Version" July 22, 2020, p6. Retreived from: https://drive.google.com/file/d/1pFYCBDFO7QP1HygYm641Npib5rWdg-CE/view.

Osinga, F. (2005) Science, Strategy and War: The Strategic Theory of John Boyd. Eburon Academic Publishers, The Netherlands. Pg. 270-279.

Rudin, Cynthia (2018). "Please Stop Explaining Black Box Models for High-Stakes Decisions", 32nd Conference on Neural Information Processing Systems (NIPS 2018), Workshop on Critiquing and Correcting Trends in Machine Learning. arXiv:1811.10154v2 5Dec2018.

Russell, Stuart and Peter Norvig (2020). Artificial Intelligence: A Modern Approach. Pearson Series in Artificial Intelligence, 4th Edition. http://aima.cs.berkeley.edu

Simaan, M and Cruz, J. B. Jr. (1975, January) Formulation of Richardson's Model of Arms Race from a Differential Game Viewpoint. *The Review of Economic Studies.* 42(1), pp. 67-77.

Smith, Ron P. (2019, November) The Influence of the Richardson Arms Race Model. Lewis Fry Richardson: His Intellectual Legacy and Influence in the Social Sciences, Pioneers in Arts, Humanities, Science,Engineering, Practice 27, p25-34, https://doi.org/10.1007/978-3-030-31589-4_3 Retrieved from https://www.researchgate.net/publication/337954743_The_Influence_of_the_Richardson_Arms_Race_Model

Sutton, Richard, Barto, Andrew (2018) Reinforcement Learning: An Introduction, 2nd Edition, pg. 1. ISBN-10: 0262039249.

Vapnik, V. (1995). The Nature of Statistical Learning. Springer, New York.

Wong, Yuna, Yurchack, John, Button, Robert W., et al., (2020) Deterrence in the Age of Thinking Machines. RAND Corporation, Santa Monica, California. ISBN: 978-1-9774-0406-0

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2020-XX