Not Oracles of the Battlefield: Safety Considerations for AI-Based Military Decision Support Systems

Emelia Probasco[1], **Matthew Burtell**[1], **Helen Toner**[1], **Tim G. J. Rudner**[2]
[1]Georgetown University
[2]New York University
{ep792, mtb122, ht429}@georgetown.edu, tim.rudner@nyu.edu

Abstract

AI-based military decision support systems that help commanders observe, orient, decide, and act on the battlefield are highly sought after by military leadership. With the advent of large language models, AI developers have begun advertising automated AI-based decision support systems designed to both analyze and act on data from the battlefield. While the desire to use decision support systems to make better decisions on the battlefield is unsurprising, the responsible deployment of such systems requires a clear understanding of the capabilities and limitations of modern machine learning models. This paper reviews recently proposed uses of AI-enables decision support systems (DSS), provides a simplified framework for considering AI-DSS capabilities and limitations, and recommends practical risk mitigations commanders might employ when operating with an AI-enabled DSS.

Introduction

Artificial intelligence promises to make sense of vast amounts of data at superhuman speeds. The military has a strong motivation to take advantage of AI, and among the most interested within the armed forces are commanders charged with making operational decisions in war. These commanders must continuously "observe, orient, decide, and act" on a fast-paced and multi-dimensional battlefield where decisions are life-or-death. The historical desire for sophisticated tools to assess and predict conditions on the battlefield to protect military forces or gain battlefield advantage has led to the creation of everything from weather modeling to campaign modeling to early warning systems. The strong, if understandable, desire for AI-enabled decision support systems (DSS) must be tempered, however, by an understanding of the capabilities and limitations of these systems, which should dictate when and how they are deployed.

In this paper, we consider the reasons why AI could help battlefield decision-making, describe some of the different systems being discussed, and offer a simplified framework for users and senior decision-makers to consider as they aim to responsibly employ AI-enabled DSS. For those less familiar with military procedures and culture, we begin with a brief history of military decisions and global approaches developed to fight through the "fog of war." We then demonstrate the widespread interest in applying AI for decision support among the world's powerful militaries. Finally, we characterize the opportunities and risks of applying AI to military decisions and offer a basic framework to guide the operationalization of these systems.

Background

Military commanders have sought support for their decisions since ancient times. Even Herodotus' Histories describe how Croesus consulted the oracles of Greece and Libya in the 6th century BCE when he decided to go to war. When he asked "if he should undertake an expedition against the Persians," the oracle assured him that "he would destroy a great empire" (Herodotus 1920). Accordingly, he went to war, and while he successfully destroyed a great empire, that empire turned out to be his own.

While oracles and AI DSS are very different, both are consulted to satisfy military commanders' desire for clarity and foresight. Commanders require awareness and desire foresight in order to make high-stakes battlefield decisions amid a cacophony of imperfect information. Senior commanders must integrate constantly changing information about their own forces—quantities, capabilities, operating status, locations, and supply chains—with similar information about their enemies while also accounting for geography, weather, time delays, and social and political context of their decisions. This information burden has long been known as the "fog of war" because it is simultaneously incomplete, confusing, complex, and continually changing.

Evidence of efforts to collect and systematically process information relevant to military decisions can be found in doctrinal memos, training books, and PowerPoints from recent conflicts. At the same time, the operations research and engineering communities have developed mathematical models that range from predicting the lethality of a specific weapon, to helping describe the lessons learned from a past battle, to predicting the evolution of a military engagement. Some of the previously developed models are narrowly focused and physics-based, such as those aimed at air-to-air engagement, weapons target assignment, or weather. Others, such as campaign models, seek to predict campaign-level outcomes and combine multiple data sources and models that are not solely based on physical limitations but rather statistical inferences

or a combination of both (Hillestad, Bennett, and Moore 1996).

Research has shown that few non-physics-based tactical or operational models have ever been adequately validated (Minguela-Castro, Heradio, and Cerrada 2022). Despite this obvious shortcoming, imperfect models are still provided to commanders to help them think through the myriad of factors that go into planning for or executing a military campaign.

Historical Military Desires Meet Modern Technologies

Despite centuries of effort, militaries have always faced limits in understanding and predicting battlefield events to make better decisions. Recently, advancements in sensors, data sources, and algorithms have revived hopes for accurate battlefield awareness and prediction. Although AI and information technology may bring significant progress, building a machine that perfectly understands and predicts war remains impossible. Military statements worldwide, as displayed in Table 1, highlight the need to distinguish between what is possible and what is desirable.

Beyond mere statements, governments have begun research projects and issued requests for proposals from military software developers for AI-enabled DSS. China's People's Liberation Army (PLA), for example, has issued calls for "battalion and company command decision-making model and human-machine teaming software." (Fedasiuk, Melot, and Murphy 2021). NATO's Science and Technology Office has initiated a research team to "investigate how reinforcement learning could be used to support decision-making." (SAS 2023). DARPA has supported the development of narrow DSS through its "In the Moment" (ITM) program, which aims to help human decision-makers in medical triage emergencies (DARPA 2024).

Globally, companies have responded to these government statements of interest, and there is now varied evidence of DSS being marketed or employed. In the United States, Palantir's AIP[1] integrates disparate data streams and applies machine learning to support commander situational awareness and target identification. Also in the United States, Scale AI's Donovan[2] is marketed as a chatbot that can be fine-tuned on sensitive or classified government data. The French multinational Thales Defense offers the AI-enabled decision support tool Anticipe, which integrates a "wargaming tool and advanced machine learning algorithms to provide military commanders with actionable insights when they need it most" (NATO 2023). Finally, China-based StarSee's "Real-time Combat Intelligence Guidance System" offers to integrate intelligence to support commander decisions (Fedasiuk, Melot, and Murphy 2021). These are just a few examples of global DSS. No two appear exactly the same, but all endeavor to make sense of the battlefield and support better battlefield decisions.

[1]See https://www.palantir.com/platforms/aip.

[2]See https://scale.com/donovan.

Types of Decision Support

The range of DSS being advertised today blurs the lines between different decision support tasks, from awareness to planning to prediction.[3] Examples include leveraging computer vision and data fusion to rapidly identify both friendly and enemy forces on a battlefield, using algorithms to spot anomalies, leveraging sentiment analysis on social media platforms, and more recently, leveraging large language models for everything from evaluating intelligence to generating courses of action (COAs). Marketing videos and military leaders have discussed applying these to artillery targeting, route planning, logistics optimization, or the apportionment of medical care in emergencies. Some proposals have even postulated that AI might be used to predict the likelihood of success of a chosen COA or the likelihood of a future conflict or political instability. Based on a review of largely American AI-enabled DSS (Appendix A contains a sample), Table 2 provides an illustrative list of tasks that have been proposed by government and industry leaders, some of which are already available. Despite AI's many capabilities, commanders must carefully distinguish between use cases for which DSS are appropriate and those deserving more skepticism. We elaborate on this distinction throughout the rest of the paper.

While the AI tasks listed in Table 2 are neatly categorized as awareness, planning and execution, and prediction, advertisements from vendors illustrate how proposed AI DSS often cross categories. For example, Scale AI, Palantir, Anduril, and Rebellion Defense each promote similar AI-enabled situational awareness tools that also support planning or predict outcomes using data fusion, computer vision (segmentation and classification systems), machine learning (anomaly detection and recommendation systems), and, most recently, generative AI and large language models.

Besides examples of expansive systems that do several types of tasks, there is also evidence of much more narrowly-scoped systems that may only address a few of the tasks listed in Table 2. For example, the Intelligence Advanced Research Projects Agency (IARPA) runs the narrowly scoped "Rapid Explanation, Analysis, and Sourcing Online (REASON)" program, which is intended to evaluate and challenge the conclusions of intelligence analysts to improve their findings and arguments (IARPA). Researchers at the Naval Research Laboratory have advertised commercial licenses for a recommender system to help military decision-makers without data analysis skills better understand and analyze data (Techlink 2024). And the company DEFCON AI markets a platform for logisticians to model supply chain disruptions and evaluate logistical resupply options (see Appendix A for more information).

These examples illustrate the types of efforts being proposed or marketed: we do not have exact information about the inner workings of each of these systems, and even if we did, the pace of progress in the field of AI would quickly banish a detailed analysis to irrelevance. Military

[3]In this analysis, decision support is different from physical control, like missile terminal guidance or swarm control for drones, to focus on those instances where a human is using an AI system to make a decision.

Country	Statement
Russia	"AI technologies can change perceptions about the size of anticipated costs and expected benefits, the balance between offensive and defensive measures, and the results of conventional and nuclear deterrence calculations, eliminating uncertainty in situation assessment, ensuring near-absolute impartiality in political and military decisions, and completely eliminating the influence of the human factor" (Shakirov 2023).
China	"Promote the modernization of emergency response management through informatization; completely upgrade monitoring and early warning capabilities, supervision, management, and law enforcement capabilities, computer-assisted command and decision-making capabilities, disaster relief fighting capabilities and social mobilization capabilities with coordination between multiple departments, and enhance international logistical supply chain service protection capabilities" (Creemers et al. 2021).
United States	"The latest advancements in data, analytics, and artificial intelligence (AI) technologies enable leaders to make better decisions faster, from the boardroom to the battlefield. Therefore, accelerating the adoption of these technologies presents an unprecedented opportunity to equip leaders at all levels of the Department with the data they need, and harness the full potential of the decision-making power of our people" (United States Department of Defense 2023).
Japan	"Accelerate decision-making through the use of Artificial Intelligence (AI)" as a target to achieve by 2027 in the Defense Buildup Plan (Japanese Ministry of Defense 2022).
Australia	"Providing high-level strategy decisions is currently beyond the state-of-the-art of AI, but there is ongoing work to broaden the applicability of AI techniques to support the humans making these complex decisions. We believe the ADF would benefit from following these developments closely and investing as appropriate. It is likely that adversaries who embrace such technology will have a dramatically reduced decision-making cycle as the capabilities in this area improve" (Moy et al. 2020).
NATO	"Improving the Alliance's situational awareness and strategic anticipation has been an important dimension of the Alliance's strengthened deterrence and defence posture. Fundamental to the Alliance's ability to shape, contest and fight is expanding knowledge and understanding, with a view to ultimately achieving cognitive superiority. This understanding shall be connected across all domains, enabled by technology, in order to maximize commanders' ability to anticipate, think, decide and act. Efforts to build better situational awareness and understanding with a view to achieving cognitive advantage over potential adversaries is a priority for the Alliance" (NATO 2021).

Table 1: Illustrative statements from global military powers.

decision-makers considering deploying DSS may face a similar dilemma—either not understanding the details of how the system works, not being up to date with the latest developments, or both. To help address this challenge, we offer a framework for commanders to consider when evaluating the appropriateness and risk of employing AI for operational decision-making.

A Simplified Framework for Military Decision Support Applications

A military leader who encounters DSS for the first time may reasonably be excited by the clarity that such systems seem to offer and the relative speed and ease by which they can integrate and process vast amounts of data. These systems can indeed add clarity and support difficult decisions. Knowing when and where they are best positioned, as well as how to mitigate their inherent risks is key to a military commander's responsible use.

The U.S. Department of Defense has established lengthy guidance for military forces to help determine the responsible application of AI (Hicks 2021). The guidance is detailed and considers many facets of AI deployment, including consideration of available compute or specific AI tools and methods, for example.[4] Similarly, past reports have gone into important details about the nuances of AI-DSS and their relationship to exercising human judgement in accordance with International Humanitarian Law (Michel 2024). While detailed guidance and nuance are important, this paper aims to serve as a simplified resource for decision-makers thinking at a high level about using or deploying AI-based DSS. The following three areas are a good starting point:

1. **Scope.** How well-defined and understood is the scope of what the system should and should not be used for?
2. **Data.** Does the training data substantiate the system's conclusions?
3. **Human-Machine Interaction.** How does the human operator understand machine outputs? What are the capabilities and limitations of the human-machine team as a single system within a given context?

To dive deeper, we consider each question in turn and then recommend mitigations where appropriate.

[4]See https://rai.tradewindai.com (Chief Data and AI Office "RAI Toolkit," Executive Summary) and https://www.diu.mil/responsible-ai-guidelines (DIU "Responsible AI Guidelines")

Situational Awareness	• Mapping friendly and enemy forces and maneuvers • Identifying anomalies in financial transactions or ship movements • Surfacing enemy investments or academic research in novel technologies • Public sentiment mapping and news media coverage and summarization • Establishing financial and social networks • Supply chain mapping and monitoring
Planning and execution	• Red teaming analytical findings • Theater re-supply planning and resiliency • Mass casualty responses • Fires/artillery target selection and weapons pairing • Course of action (COA) generation and red-teaming • Command and control of remote units
Prediction	• Predictive maintenance • Early warning of preparations for war • Forecasting political instability or mass movements • Forecasting combat outcomes or COA success probabilities

Table 2: Illustrative list of AI applications to military decisions.

Scope

How well-defined and understood is the scope of what the system should and should not be used for?

Some of the DSS available today are very tightly scoped, such as an AI tool that can use an image to identify "viewsheds," i.e., areas that an observer at a given point (such as a sniper) can or cannot see (Doyle 2013). The clearer and more well-defined the scope is for a given AI tool, the easier it is to ensure that it works as intended when used within that scope, and the less likely it is to be used in ways it was not tested and validated for. When considering whether the scope of a given DSS is appropriate, commanders should be on the lookout for:

Context shifts. A general challenge with using modern AI systems in high-stakes settings is that they are prone to fail if used in settings that are meaningfully different from the settings they were trained on (Rudner and Toner 2021a,b, 2024). For example, models previously successful at predicting shopping trends, traffic, and supply chains began to fail in 2020—as the COVID pandemic spread and habits changed (Steinhardt and Toner 2020). Similarly, it would be fraught to rely on sentiment analysis algorithms optimized for only one dialect of Arabic when trying to assess social movements in the Middle East (El-Masri, Altrabsheh, and Mansour 2017; Medhat, Hassan, and Korashy 2014). In some instances, human operators can easily observe degradations in AI performance, but this is not always the case. Figure 3 depicts an example of a potential "distribution shift," drawn from research on identifying flooded buildings from overhead imagery (Rudner et al. 2019). A model trained on data from Houston would not necessarily perform well on data from Russia or Germany, where the landscapes in question are very different. Systems should be carefully tested under different operational conditions to ensure that the scope of situations where they perform well vs. poorly is known and communicated to users.

Figure 1: Distribution Shifts in Image Processing: Overhead image of Houston, Russia, and Germany (Rudner et al. 2019).

Projection and prediction. Some decision support tools propose to ingest large data sets not to provide awareness but rather to predict the future, such as the possibility of social uprisings or military attacks. On the one hand, these sorts of predictions are not new: humans have long forecasted the weather, and today deliver highly-accurate near-term predictions. But weather is a phenomenon that has been long observed, is well instrumented, and follows reasonably well-understood physical laws—and it is still reliably predictable only a week or so ahead of time. Human interactions and decisions are far more complex, less understood, and not fully observable (see next section on data). DSS whose scope includes making projections or predictions should be subject to additional scrutiny unless those predictions are based on well-understood physical laws and anchored in directly applicable data (e.g., predicting impact points based on ballistic missile trajectories).

Flexible or unclearly scoped systems. The advent of large language models (LLMs) has not gone unnoticed among companies developing military DSS. Companies including Scale AI and Palantir have advertised DSS that use the general-

purpose flexibility that LLMs offer.[5] These demos are impressive but raise several concerns, both regarding human-machine interaction (see below) and scope.

There is nothing inherently inappropriate about the idea of applying LLMs to decision support applications. However, the importance of careful testing and validation means that any use of LLMs—like other uses of AI for decision support—should be carefully scoped to a defined set of tasks. Some early demos of LLM-based products, in contrast, purport to be something more like an all-purpose "battle buddy" that can assist the user with tasks as wide-ranging as identifying enemy units, searching the literature for contextual facts about combatant countries, analyzing intelligence, generating courses of action, and directly sending digital commands to an autonomous or remotely piloted vehicle. The breadth of these tasks, combined with the natural-language interface that encourages the operator to think of the system as akin to a human companion, is cause for concern. Without clear boundaries around which tasks have been thoroughly tested and validated, operators are likely to stretch the limits of such systems, pushing them to assist with tasks that go beyond the contexts and purposes for which they were developed and validated.

Irreducible uncertainty. The Russian statement in Table 1 refers to the possibility of "eliminating uncertainty." This phrase demonstrates a belief not limited to Russia—that it is possible, in principle, to reduce uncertainty to zero. In most real-world settings, this is not possible. Consider, for example, a fair coin flip. You could flip a coin repeatedly and—with sufficiently many coin tosses—would be able to become highly certain that the probability of a flip landing heads is 50%. However, this would not help reduce any uncertainty about whether the next flip will land heads or tails. In other words, the uncertainty is irreducible. The users of AI DSS may want to know things like "Where will the enemies move?" or the real-world equivalent of "What move will my opponent play in rock, paper, scissors?" Questions like these have an inherent level of uncertainty that cannot be fully reduced, exacerbated by the near-impossibility of taking perfect measurements and modeling complex systems. In sum, this means that humans must still exercise judgement in making battlefield decisions.

Data

Does the training data substantiate the system's conclusions? How is the data collected and how frequently is it updated? What abnormalities, outliers, or irregularities might be present in the data? How was the data generated?

Data quality and fidelity. Modern machine learning systems for computer vision and natural language processing tasks are developed using large quantities of data. When high-quality, relevant data is plentiful, building a system that works well is easier. Weather prediction is an example: We have enormous quantities of high-fidelity data on exactly this kind of phenomenon—temperature, air pressure, cloud formation, precipitation—and a strong mechanistic understanding of the physical dynamics that drive weather patterns. Importantly, the data is collected in the same environment we are trying to predict. However, even under ideal circumstances, the complexity of weather events can lead to prediction errors—especially over long prediction horizons.

By contrast, although human behavior may follow somewhat predictable patterns, human-based data, such as opinions and thoughts, are only observable indirectly by a person's actions or words and do not follow physical laws. Further, data collected in one group may not be predictive for other groups or demographics. Researchers sometimes try to augment their datasets using simulated data, but the effectiveness of this approach depends on the quality of the simulations employed. Simulated data works only insofar as it is an accurate representation of reality. This is much easier to do when we understand the mechanics involved and can validate the simulations through testing, as in the case of physics-based systems like missiles, tanks, planes, and ships. Simulation is far harder to do when we don't understand or cannot test, such as in social decision-making. In cases where we lack a clear comprehension of the mechanisms that connect inputs to outputs—as is true regarding almost any social or political question—simulated data may not be effective.

Skewed data. Military commanders struggle to obtain accurate data about their own forces, let alone information about the enemy. For friendly forces, some units in the field can have poor communications and may not accurately transmit data about their position or status. Other platforms may flood the data stream with a single sensor in a particular area, causing humans and computers to over-index on the single source. When it comes to adversaries, some are more adept at avoiding intelligence collection than others, some may be adept at active deception, and finally available data on some adversaries may be scarcer than others. While some information is more useful than no information, and complete information may be impossible if not harmfully slow to assemble, this biased data can skew a DSS in a way that must be communicated to a user.[6] Furthermore, commanders should be especially wary of how biases in AI-enabled DSS might be amplified when they align with the personal or cultural biases of operators (Michel 2024).

Human-based data presents different issues. Social media platforms often reflect a number of human biases—demographics may vary from one social media platform to another, and discourse on social media, which is shaped by technical, psychological, or social factors, may not reflect real-world behaviors or even opinions. Thus, using AI to predict violent uprisings on the basis of open-source infor-

[5]For demonstration videos, see https://www.youtube.com/watch?v=XEM5qz_HOU (Palantir AIP) and https://www.youtube.com/watch?v=0aBxvzdoyow (Scale Donovan).

[6]General Colin Powell made famous the 40/70 rule, implying that 40% was the minimum amount of information a military commander needed to make a decision about a situation and 70% of the information was the maximum, as decisions made after 70% of the information was gathered were prone to be made too late to have an effect in time sensitive scenarios.

mation about the sentiment of the populace is an example of a problem where the underlying data is far weaker and the application of AI is suspect.

Scarce data While it is possible to gather large quantities of data on social media patterns, news reports, and market movements, the number of instances of the outcome to be predicted—uprisings for example—can be extremely small (hundreds at most, compared with orders of magnitude more examples of weather events). Along the same lines, the military often faces a small data problem as the adversary is actively trying to conceal its most valuable (and often rare) capabilities. In the absence of adequate observations, AI-based DSS are less valuable than traditional modes of intelligence analysis, which can combine insights and inferences that rely on a richer understanding of the relevant context in any particular case.

Human-System Integration

How does the human operator understand machine outputs? What are the capabilities and limitations of the human-machine team as a single system within a given context?

When evaluating whether it is appropriate to use an AI system in a decision support context, it is crucial to consider the properties of the AI system itself and also how human operators will interact with it. Three facets of human-system integration can pose important risks to the responsible use of DSS:

Setting false expectations with LLMs. LLMs present distinct challenges from a human factors perspective. Notably, their propensity for confidently asserting factually incorrect information, often referred to as "hallucinations" or "confabulations," can be highly misleading for users. Beyond hallucinations, LLMs exhibit further problematic behaviors, such as fabricating justifications for recommendations. These fabricated rationales tend to align with user expectations rather than reflecting the true underlying reasoning process (Turpin et al. 2023). Research has unfortunately demonstrated that such unfaithful explanations can increase user acceptance of erroneous recommendations (Bansal et al. 2023).

While AI systems can provide estimates of event probabilities, an inherent and irreducible uncertainty regarding future outcomes remains. The distinction between a well-defined DSS and one venturing into the riskier domain of prediction often hinges on communication. For instance, a system could display "there is a 90% chance of an uprising," constituting a prediction, or it could present "in 90% of cases where indicators X, Y, and Z were observed, social unrest was documented in the subsequent N weeks," offering awareness. The latter fosters a more nuanced understanding of potential risks and avoids the pitfalls of definitive pronouncements.

Human biases. Untrained users can be subject to automation bias, assuming that an automated system knows best even when their own judgment says otherwise. This can be especially problematic in situations that involve the integration of large amounts of data, where computers are seen as superior. The potential for humans to defer to algorithmic recommendations is well-documented and could be particularly problematic in a situation as dynamic, uncertain, and consequential as warfare (Kahn 2024).

Organizational biases. DSS can speed up decision-making and reduce the personnel required for certain tasks, but organizational assumptions about speed and human resources can lead to hasty or under-resourced decision-making and put pressure on military commanders to "do more with less." While some decisions may become faster as a consequence of AI and autonomous systems, the speed and quantity of decisions should be calibrated to the quality of the decision-making processes and tools as well as contextual risk. For example, reports of Israel's Lavender and Gospel targeting indicate that before October 2023 the system was used as a decision support aid with extensive oversight, whereas afterwards organizational policy shifted and "We didn't go through them one by one—we put everything into automated systems" (Iraqi 2024). Relatedly, the apparent ease of using some DSS may predispose organizations and the people who work within them to use these systems in all cases, when in fact they may be best suited to only the most extreme situations when risk tolerance is high.

Recommendations

DSS have strategic benefits in certain contexts, but those benefits are not without risks which military commanders should work to control. With due consideration for current AI limitations and some of the guidelines already offered by, for example, the Department of Defense's Responsible AI Toolkit, we offer the following recommendations to leaders evaluating how to deploy these systems and mitigate their risks. We note that each of these recommendations has to do with humans, not with the technologies themselves.

Technical risk mitigation is important and there are myriad efforts underway to technically improve AI systems. Here, however, we take as a premise that human governance actions will remain essential to mitigating current and future shortfalls of AI-DSS and, moreover, that humans are the last line of defense for ensuring AI-DSS employment is done in compliance with international humanitarian law.

1. **Establish context- and risk-based criteria to govern use**

   A commander deploying a DSS should take into consideration the time, place, and context of its application. Furthermore, deployment of a DSS should be reversible, and deployments should be adjusted or ended as factors on the ground change in ways that would affect AI performance. Commanders should establish a process and criteria for the continued deployment of AI-based DSS. That process must take into account not only the scope of the technology and the operational context but also the readiness of the organization to properly leverage the DSS through codified and tailored tactics, techniques, and procedures that will keep its operations within the rules

of engagement. In many ways this approach is similar to how the military already approaches rules of engagement and weapons postures—adjusting the authorization for operators to activate certain systems or employ processes around those systems depending on the tactical or strategic context.

2. **Train and qualify AI-enabled DSS operators**
   Any operator tasked with using an AI-based DSS, especially one supporting targeting or the employment of weapons. That knowledge starts with study and experimentation in safe contexts but must build quickly from that base. A qualification regime for a DSS, especially those guiding operational decisions or weapons employment, should have both schoolhouse and experiential components. Those DSS involved in lethal decisions could also be paired with rigorous, tailored examinations that lead to an official qualification designation for a DSS operator by a unit commander for operators that is commensurate with their role in relation to the AI DSS. The National Geospatial Agency's GEOINT Responsible AI Training (GREAT) Program is an early leader in designing and implementing a process for qualification and certification for users (from personal correspondence with Dr. Anna Rubinstein).
3. **Establish a continuous unit certification cycle**
   Even with thoughtful design and user qualification, leaders should certify an organization's ability to faithfully execute a commander's intent using an AI-enabled DSS. Because the environment in which a DSS is deployed is constantly changing, assessments should be frequent if not continuous. Since DSS are also repositories for data, logging and retaining data on the use and effectiveness of DSS should be routine. Commanders should consider how to embed or share DSS system performance metrics with trained data scientists, experts in AI evaluation, and/or operations analysts both to validate the continued use of the system and to evaluate the effectiveness of a unit leveraging a DSS.
4. **Designate of a unit AI safety officer**
   AI risks and opportunities are rapidly changing and show no signs of slowing. Moreover, there is a pressing need to establish a broad-based AI literacy to help avoid some of the DSS risks we detail above. While there are unique factors to this challenge, there are similarities to other safety risks the military has addressed in the past. Depending on the mission, for example, military units often have Occupational Safety, Weapons Safety, Test Safety, or other established programs. AI safety officers could serve as a local, educated resource for operators as well as a conduit for sharing new information from central organizations like the Chief Digital and Artificial Intelligence Office (CDAO).
5. **Document incidents and harms from AI systems to support knowledge sharing with developers and other users**
   AI systems have flaws and human users will make mistakes. To best avoid repeating past mistakes, the Responsible AI Toolkit has made clear that AI harms should be documented. As of publication, CDAO is still developing a process specific to the military's needs, though several methods of reporting have been proposed for general use. We support these efforts as they help to avoid past mistakes and will build trust through transparency. We would suggest that AI safety officers take on the responsibility for documenting these harms in the same way that safety officers are often responsible for mishap reporting. Moreover, while military leaders may be predisposed to classifying or otherwise keeping all reports of incidents within DOD official channels only, they should consider the value of sharing incident reports. By making at least some reports public, the military may help prevent inadvertent harm in other militaries or in civilian contexts. It could also build trust with the public through such transparency efforts. Moreover, making incident reporting a norm might help enable clearer communications between competitors in the event an AI system fails in a way that might be misinterpreted as aggressive.

Conclusions

AI may improve the quality and speed of decision-making on the battlefield, but it cannot replace human judgment. The temptation to believe that AI DSS are all-knowing is strong: They can draw on disparate data sources, integrate vast amounts of information, and generate recommendations at superhuman speeds. Moreover, their capabilities seem almost purpose fit for the unique challenges of war and the fog of battle. But AI DSS also have critical weaknesses that require acknowledgment and intervention and operators must beware of any automation bias they might harbor.

To seize the opportunity of AI DSS in battle, commanders must prepare themselves and their teams to use them correctly. Considering questions of scope, data, and human-machine integration are important starting points for avoiding the weaknesses of these systems. These weaknesses alone are not necessarily reasons to avoid AI DSS altogether, but they must be paired thoughtfully with human intelligence and judgment to achieve the best outcomes.

A: Further Details on DSS Projects

Organization	Project Name	Statement
Anduril	Lattice	"Lattice streamlines the complexity of the decision-making process by presenting decision points – not noise – and using deep learning models to present recommended decision support to operators. Lattice enables real-time command and control over manned and unmanned assets across multiple domains, distributed geographies, and in contested communications environments."
Clarifai	Multiple projects	"Rapidly turn mountains of data into plans of action for decision advantage to support the warfighter and the supply chain."
DARPA	In The Moment	"The Defense Sciences Office (DSO) at the Defense Advanced Research Projects Agency (DARPA) is soliciting innovative research proposals for research and technology development that supports the building, evaluating, and fielding of algorithmic decision-makers that can assume human-off-the-loop decision-making responsibilities in difficult domains, such as combat medical triage."
DEFCON AI	DEFCON AI	"DEFCON AI, a next-generation software company building the foundation to enhance the modeling, simulation, and artificial intelligence enterprise across the Department of Defense (DoD), announced that it closed an [Air Force contract] to accelerate the transition from prototype to production code for DEFCON AI's operational-level logistics and mobility training software."
IARPA	REASON	"The Rapid Explanation, Analysis, and Sourcing Online (REASON) Program aims to develop technology that will enable intelligence analysts to substantially increase the quality of argumentation in their analytic reports through more effective use of evidence and reasoning."
Johns Hopkins APL	Wolf Howl	"We'll give commanders the ability to 'wargame' different strategies from mission to unit within a given time frame or risk tolerance. That way, humans and computing machines can focus on aspects of planning they are each currently better suited to, and you really try to get the best of both worlds."
Leidos	Global Planning and Monitoring (GLIMPS)	"The GLIMPS provides accurate, global forecasts on defined lead times of up to five years, focused on turbulent and complex environments, in order to provide the information needed to adapt to changing needs and resources. GLIMPS technology forecasts the effects of poverty, environmental degradation, political instability, and social tensions through big-data mining and machine learning of millions of open-source intelligence data points to discover the unseen relationships between indicators of stress and locations of potential instability."
NATO Science and Technology	ANTICIPE	"ANTICIPE is designed to aid decision-making in an operational setting, using a built-in wargaming tool and advanced machine learning algorithms."
Palantir	AIP	"By leveraging the latent power of organizational data alongside interfaces for intelligent, fast decision making, AIP provides next-generation tooling."
Rebellion Defense	Iris	"Iris leverages state-of-the-art AI trajectory prediction methods to quickly find high-interest entities among cluttered environments for further investigation."
Scale AI	Donovan	"AI-powered decision-making for defense."

Table 3: DSS Project Details. We considered a number of AI DSS that are being developed, advertised, requested, or used today. While the list is not all-inclusive, it is intended to provide a representative snapshot of the systems that have been requested, proposed, or developed.

Acknowledgments

For feedback and assistance, we would like to thank John Bansemer, Sam Bresnick, Lauren Lassiter, Igor Mikolic-Torreira, and Michael O'Conner.

References

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. S. 2023. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance.

Creemers, R.; Dorwart, H.; Neville, K.; Schaefer, K.; Costigan, J.; and Webster, G. 2021. Translation: 14th Five-Year Plan for National Informatization (Dec 2021).

DARPA. 2024. In the Moment. https://www.darpa.mil/program/in-the-moment. Accessed: 2024-07-27.

Doyle, R. J. J. 2013. Real-time lines-of-sight and viewsheds determination system. U.S. Patent 8,400,448. Issued March 19, 2013.

El-Masri, M.; Altrabsheh, N.; and Mansour, H. 2017. Successes and Challenges of Arabic Sentiment Analysis Research: A Literature Review. *Social Network Analysis and Mining*, 7(1): 54.

Fedasiuk, R.; Melot, J.; and Murphy, B. 2021. Harnessed Lightning: How the Chinese Military is Adopting Artificial Intelligence. Technical report, Center for Security and Emerging Technology, Washington, DC. Accessed: 2024-07-27.

Herodotus. 1920. *Histories*. Cambridge, MA: Harvard University Press.

Hicks, K. 2021. Implementing Responsible Artificial Intelligence in the Department of Defense. https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPONSIBLEARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF. Accessed: 2024-07-27.

Hillestad, R.; Bennett, B. E.; and Moore, L. R. 1996. Modeling for Campaign Analysis: Lessons for the Next Generation of Models: Executive Summary.

(IARPA), I. A. R. P. A. 2024. Reason: Rapid Explanation, Analysis and Sourcing Online. Accessed: 2024-03-12.

Iraqi, A. 2024. 'Lavender': The AI Machine Directing Israel's Bombing Spree in Gaza. *+972 Magazine*.

Japanese Ministry of Defense. 2022. Defense Buildup Program. https://www.mod.go.jp/j/policy/agenda/guideline/plan/pdf/program_en.pdf. Accessed: 2024-07-27.

Kahn, L. 2024. Automation Bias. Forthcoming.

Medhat, W.; Hassan, A.; and Korashy, H. 2014. Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Engineering Journal*, 5(4): 1093–1113.

Michel, A. H. 2024. Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making. Technical report, International Committee of the Red Cross (ICRC). Prepared for the International Committee of the Red Cross (ICRC).

Minguela-Castro, G.; Heradio, R.; and Cerrada, C. 2022. Automated Support for Battle Decision Making: A Systematic Literature Review. *Military Operations Research Society*, 27(4): 5–24.

Moy, G.; Shekh, S.; Oxenham, M.; and Ellis-Steinborner, S. 2020. Recent Advances in Artificial Intelligence and their Impact on Defence. Technical Report 17, Defence Science and Technology Group, Commonwealth of Australia.

NATO. 2021. NATO Warfighting Capstone Concept. https://www.act.nato.int/wp-content/uploads/2023/06/NWCC-Glossy-18-MAY.pdf. Accessed: 2024-07-27.

NATO. 2023. NATO and AI. https://www.linkedin.com/posts/natosto_nato-ai-activity-7178036772948340736-I4zD/. Accessed: 2024-07-27.

Rudner, T. G. J.; Rußwurm, M.; Fil, J.; Pelich, R.; Bischke, B.; Kopackova, V.; and Bilinski, P. 2019. Multi³Net: Segmenting Flooded Buildings via Fusion of Multiresolution, Multisensor, and Multitemporal Satellite Imagery. In *Proceedings of the Thirty-Three AAAI Conference on Artificial Intelligence*.

Rudner, T. G. J.; and Toner, H. 2021a. Key Concepts in AI Safety: An Overview. In *CSET Issue Briefs*.

Rudner, T. G. J.; and Toner, H. 2021b. Key Concepts in AI Safety: Robustness and Adversarial Examples. In *CSET Issue Briefs*.

Rudner, T. G. J.; and Toner, H. 2024. Key Concepts in AI Safety: Reliable Uncertainty Quantification in Machine Learning. In *CSET Issue Briefs*.

SAS. 2023. NATO STO research team explores ways to support decision making with AI. Accessed: 2024-07-27.

Shakirov, O. 2023. Russian thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making. https://www.europeanleadershipnetwork.org/wp-content/uploads/2023/11/Russian-bibliography.pdf. Accessed: 2024-07-27.

Steinhardt, J.; and Toner, H. 2020. Why Robustness Is Key to Deploying AI. https://www.brookings.edu/articles/why-robustness-is-key-to-deploying-ai/. Accessed: 2024-04-25.

Techlink. 2024. Recommender system for enhancing exploratory data analysis. https://techlinkcenter.org/technologies/real-time-3d-viewsheds-imaging-software/7088ca39-72cd-432e-b4db-3a35545db3b1. Accessed: 2024-07-27.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

United States Department of Defense. 2023. DoD Data, Analytics, and Artificial Intelligence Adoption Strategy. https://media.defense.gov/2023/Nov/02/2003333300/-1/-1/1/DOD_DATA_ANALYTICS_AI_ADOPTION_STRATEGY.pdf. Accessed: 2024-07-27.