



# Recent Insights in Responsible AI Development and Deployment in National Defense: A Review of Literature, 2022–2024

Ryan Jenkins<sup>a</sup>, John P. Sullins<sup>b</sup>, Obinna Kalu<sup>b</sup>, Aisha Kamath<sup>b</sup> and Kritтана Phumjam<sup>b</sup>

<sup>a</sup>Philosophy Department, Cal Poly (California Polytechnic State University), San Luis Obispo, California, USA;

<sup>b</sup>Department of Philosophy, Sonoma State University, Rohnert Park, California, USA

## ABSTRACT

This “literature refresh” identifies the most relevant new research in AI and robotic systems ethics from January 1, 2022 to January 31, 2024. Our selection methodology consisted of traditional research methods as well as novel human-AI teaming techniques, leveraging the expert human judgment of the authors, enhanced with a collection of AI and computational tools. We have identified stable trends in the critiques of the use of AI in the defense and security domain that cluster around worries about machine bias as well as the propensity of the technology to exacerbate human cognitive biases. Training data ambiguities, irregularities or untrustworthy data, and outright hacking of training sets are notable problems reported by the papers in our research set. This limits the trustworthiness of some systems which is heightened by the “black box” nature of many of these technologies which makes accountability and testing difficult. Given the speed and immense scale of operations that AI systems are involved in, there is a pronounced drift away from the reliance on “human in the loop” and “human on the loop” as the gold standard. We are now at the stage where a new ethical paradigm or solution is needed.

## KEYWORDS

Responsible artificial intelligence; autonomous weapons; machine bias; sociotechnical systems; meaningful human control

## Introduction

As part of an ongoing project to create a risk management framework for the use of AI in defense and national security applications, we have conducted a “literature refresh” on relevant papers published in 2022 and later. Here, we discuss our methodology and present an extensive survey of issues that could arise in the machine learning pipeline for systems in defense and national security.

As with any literature review, at best we can aspire to provide a snapshot of an ongoing conversation – this literature is active, as the hundreds of papers uncovered in the last few years show. We term this survey a “literature refresh” because we are stepping into a debate that is substantive and dynamic. A flurry of recent work has appeared, especially since the publication of the United States Department of Defense’s “Responsible AI

**CONTACT** Ryan Jenkins ✉ ryjenkin@calpoly.edu 📧 Philosophy Department, Cal Poly (California Polytechnic State University), San Luis Obispo, California, USA

© 2025 Informa UK Limited, trading as Taylor & Francis Group

Principles” in 2020 and its ongoing work to guide the implementation of those principles. We aim to summarize this more recent work without rehashing decades-old debates.

Moreover, there will surely be relevant papers published even as this paper is published. Let us designate, therefore, January 31, 2024 as the cutoff date for our literature survey, as this is the publication date of the most recent article we surfaced with our keyword search and considered for inclusion (Grodzinsky, Wolf, and Miller 2024). This entails that many worthy papers will be excluded simply because they were published more recently.<sup>1</sup> While unfortunate, this is inevitable.

## Methodology

Our goal was to survey the literature, starting in 2022, to identify potential risks from the use of AI-enabled systems in defense and national security. To surface articles for review, we generated key phrases (Table 1) to use with several databases (Table 2), including through our home institutions’ library searches which consolidate access to several databases.

The articles that were returned were imported into Zotero and automatically de-duplicated. After deduplication, we were left with 278 articles. We manually tagged 130 article abstracts according to the scheme: 1-relevant-high, 2-relevant-low, and 3-not-relevant. After this initial round of manual tagging, we fine-tuned a custom AI model, following the OpenAI best practices found in OpenAI’s documentation<sup>2</sup>, using base model gpt-3.5-turbo-0125 over three training epochs, with an ultimate training loss of 0.0000 [*sic*]. Our fine-tuned GPT tagged the remainder of the abstracts. After this process, we were left with 104 articles published since 2022 and tagged 1-relevant-high. A full bibliography of articles tagged 1-relevant-high is available at **Appendix A**. The results at each stage of this process are detailed in **Figure 1**, a modified PRISMA diagram which incorporates the contributions of the fine-tuned AI model.

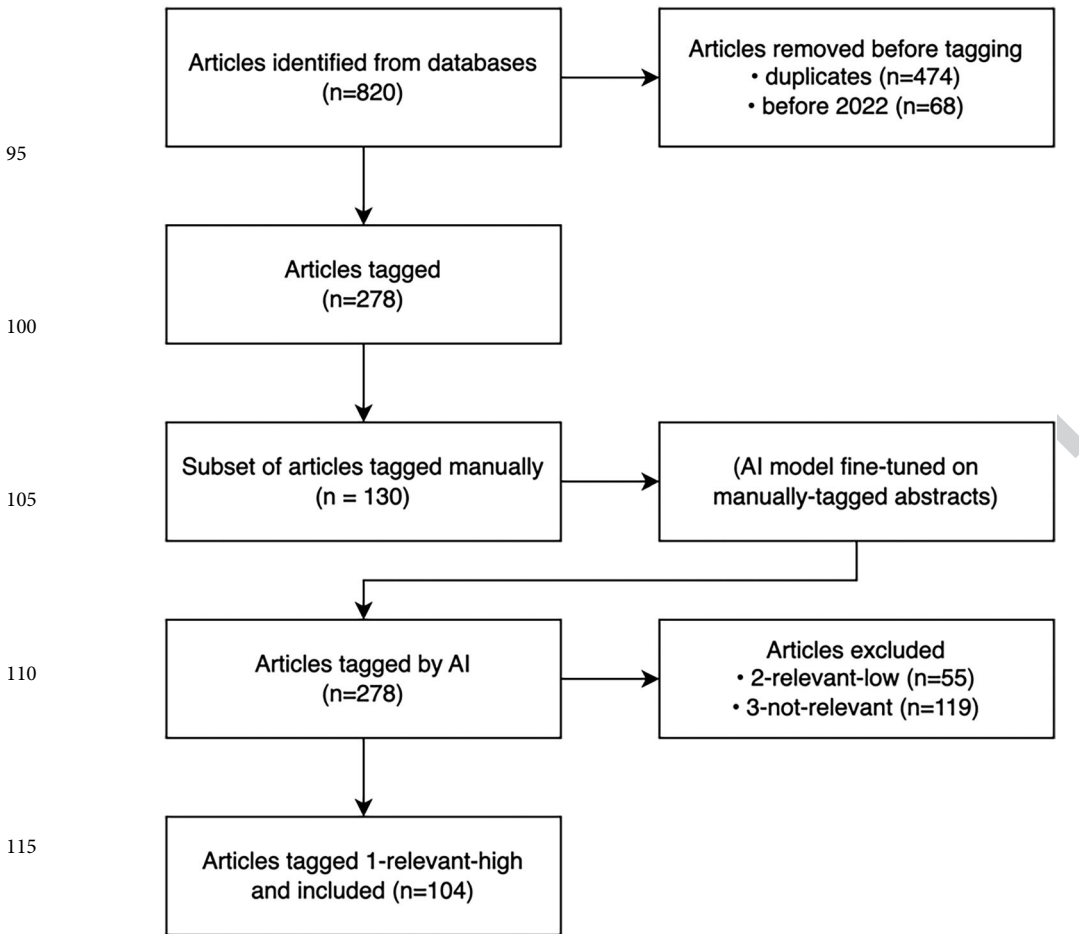
After this, we systematically applied a list of questions developed by the Machine Learning Impact Initiative (MLII) at Northwestern University to the collected literature.

**Table 1.** Key phrases used in literature search.

AI Risk Management in Defense	AI Operational Challenges in Military
Autonomous Systems Risk Assessment	Autonomous Robotics in Security
AI Deployment in National Security	AI Data Security in Military Operations
AI and Military Decision Making	AI and Human Control in Warfare
AI System Vulnerabilities in Defense	AI Surveillance and Reconnaissance Risks
Malicious Use of AI in Security	AI Sociotechnical Risks in Defense
AI Risk Mitigation Strategies	AI Risk and Mitigation in National Defense
AI Systemic Risks in Armed Forces	

**Table 2.** Databases used for literature refresh.

Google Scholar	arXiv.org
Elicit.com	EBSCOhost
IEEE Xplore	Elsevier ScienceDirect
PubMed Central	Academic Search Premier
Web of Science	IEEE Electronic Library Conference Proceedings
Springer Books	Thomson Reuters Westlaw
Springer Journals	Semantic Scholar
	ProQuest



**Figure 1.** Databases used for literature refresh.

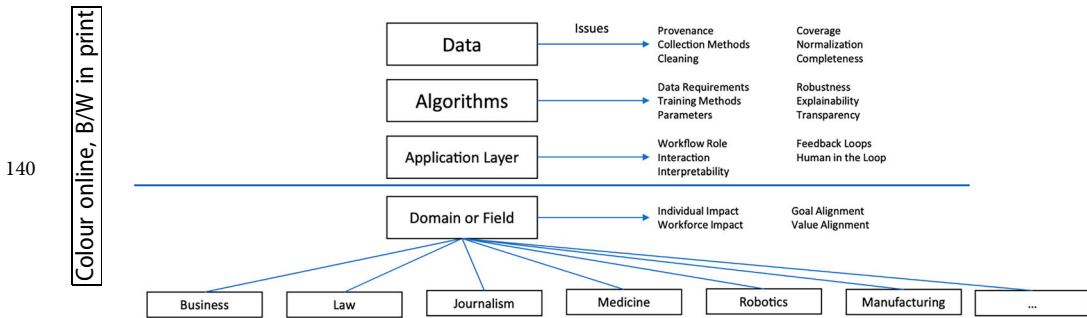
120

125

130

135

The MLII was a collaboration between Underwriters Laboratories and the McCormick School of Engineering at Northwestern University to develop a relatively exhaustive, domain-agnostic framework for surfacing relevant considerations at each stage of the machine learning (ML) development and deployment pipeline, covering the phases of data procurement and preparation, algorithm choice and model training, user interaction and framing, and finally domain-specific evaluation. Evaluating each step in the pipeline in light of domain-specific goals and values allows us to cast technical decisions in the light of ethically significant impacts. [Figure 2](#), taken from Jenkins et al. (2023), illustrates this decomposition of the ML development and deployment pipeline. A complete list of the MLII questions that were queried over our collected literature is included in [Appendix B](#). These questions were asked of the collected PDFs using Python module paper-qa, which builds an index of vector embeddings of PDFs and queries them using retrieval-augmented generation (RAG), assembling an answer from contextually relevant excerpts, using a mixture of GPT-3.5 and GPT 4 (Lála et al. 2023). These answers were compiled into a manuscript. At this point, our human reviewers, Drs. Jenkins and Sullins with assistance from Ms. Kamath, reviewed and revised these answers.



**Figure 2.** Machine Learning Impact Initiative (MLII) framework for the evaluation of the impact of systems based on machine learning (Jenkins et al. 2023, Figure 1).

Our process should be viewed as a conscious experiment in human-AI teaming, where we leverage both expert human judgment and a collection of AI tools to broaden, filter, categorize, query, and summarize the discovered literature. By casting a wide, AI-enabled search net, we aim to encompass all relevant sources and their citations, ensuring comprehensive coverage of the field while retaining human expertise in the loop. Despite the extensive use of AI, human oversight remains crucial, especially in abstract tagging, in the collating and revision process, and in the final evaluation of the discovered risks and mitigations. This methodology not only enhances the efficiency and breadth of our review but also allows us to explore the dynamics of human-AI collaboration in the context of academic research.<sup>3</sup>

## MLII questions review<sup>4</sup>

### *Data collection and preprocessing*

The machine learning pipeline begins with collecting and processing data to train a model on a task of prediction, classification, and so on. In this step, using inaccurate or biased data can lead to risks that significantly impact decision-making and operational effectiveness. These risks include incorrect results (i.e. false positives or negatives), misidentification of targets, and the provision of misleading information (Rashid 2023). Such inaccuracies in turn can result in flawed strategic decisions, potentially leading to unintended escalation and affecting international strategic stability (Rashid 2023).

Furthermore, data that demonstrates systematic biases can skew AI-based evaluations, potentially leading to severe consequences like the misinterpretation of tactical alerts (Rashid 2023). Relying on flawed data can perpetuate existing biases, a phenomenon that has been demonstrated extensively in non-military domains, underscoring the potential for incorrect predictions and strategies that could endanger operational effectiveness and safety (Goldfarb and Lindsay 2022). Incomplete or biased data can lead to AI systems developing flawed strategies, as seen in Amazon’s hiring AI and chess AI examples, where the lack of successful women applicants and training on Grandmaster games, respectively, led to biased and flawed decision-making (Goldfarb and Lindsay 2022). Amazon’s hiring AI inadvertently learned to screen out resumes containing the word “women” (Goldfarb and Lindsay 2022).

185 Frustratingly, human biases in selecting data can limit decision-making, contrasting with AI's ability to consider all available data without bias (Davis 2022). Biases, outdated information, and lack of representativeness in datasets must be adequately addressed (Carter 2022). This species of bias not only compromises the decision-making process but can also lead to severe consequences in military operations, including fratricide, civilian casualties, and unjust targeting of individuals (Goldfarb and Lindsay 2022; Limata 2023). These biases can result in unfair treatment of specific persons or groups by failing to accurately represent diverse populations or evaluate data appropriately, thus compromising decision-making accuracy and operational integrity (Limata 2023; Steimers and Schneider 2022). Furthermore, the reliance on biased AI systems can undermine mission effectiveness and erode trust in autonomous systems, potentially violating key principles of armed conflict such as discrimination and proportionality (Limata 2023).

190 The relevance and completeness of features captured by data are crucial for the effectiveness of AI systems. Relevant and complete data ensure that AI models are trained on accurate, precise, and representative datasets, which is essential for model generalization and achieving optimal performance (Steimers and Schneider 2022). Overlooking critical features in data can result in inaccurate models and potentially unsafe AI system applications, highlighting the importance of capturing the complexity of the intended task and usage environment (Steimers and Schneider 2022). Furthermore, neglecting critical data features can impact the safety, privacy, and functionality of systems, underscoring the necessity of designing robust systems that can handle uncertainties effectively (Santoso and Finn 2023). Ensuring data relevance and completeness is fundamental to mitigating risks and enhancing the reliability and safety of AI systems.

200 Similarly, ambiguous concepts and ontologies in AI training data can result in models memorizing and disclosing sensitive details, posing risks of data privacy breaches and leading to unintended outcomes (Blauth, Gstrein, and Zwitter 2022). The complexity of tasks and operational environments, especially, for example, real-world contexts and missions, can lead to incomplete specifications in AI systems, exacerbating the risk of misinterpretations and decision-making errors in complex environments (Steimers and Schneider 2022). Furthermore, these oversights can contribute to severe societal, economic, and ethically concerning impacts, including manipulation of information, economic disruption, and social inequality (Wirtz, Weyerer, and Kehl 2022). Overall, meticulous attention to feature selection and dataset auditing is crucial for preventing misinterpretations, ensuring the reliability and effectiveness of AI systems, and mitigating potential risks associated with biased outcomes, compromised privacy, and lack of trust in AI-based decisions (Wirtz, Weyerer, and Kehl 2022).

210 The methods used to collect data for AI systems can have significant security implications, potentially compromising the integrity and confidentiality of sensitive information. "Attribute inference attacks," where attackers use publicly available data to infer private information, highlight the vulnerability of data collection methods to privacy breaches (Sangwan, Badr, and Srinivasan 2023). The possibility of collating, for example, unclassified information to reveal insights that would be classified or controlled unclassified (CUI) is also worrisome. Research should be conducted into how to avoid disclosing information which, while benign in isolation, could be fused or combined to compromise national security.

Data collection methods must ensure the integrity, availability, and confidentiality of data, as adversaries may seek to steal, manipulate, or deny access to AI learning data, posing a threat to cybersecurity within the military enterprise (Goldfarb and Lindsay 2022). The integration of public data, even when anonymized, increases the risk of identification, underscoring the challenges in maintaining data privacy (Carter 2022). Moreover, the collection process itself can introduce biases and anomalies, further compromising data integrity (Goldfarb and Lindsay 2022).

AI systems are susceptible to various cyberattacks during their training and testing stages, such as “data poisoning” attacks, which can manipulate training data or present falsified data, thereby threatening the integrity and confidentiality of the information processed by these systems (Sangwan, Badr, and Srinivasan 2023). False information can be ingested during the training process, thereby undermining trustworthiness and causing potential financial, privacy, or national security impacts (Blauth, Gstrein, and Zwitter 2022). These vulnerabilities call for the implementation of robust security measures, including AI systems designed to be secure from the outset, to mitigate potential threats (Sangwan, Badr, and Srinivasan 2023). Solutions such as blockchain-based data storage have been proposed to address these vulnerabilities (Sangwan, Badr, and Srinivasan 2023).

This underscores the importance of cleaning, curating, and consistently integrating data sources to maintain the integrity, accessibility, and accuracy of AI learning, thereby mitigating risks in decision-making processes (Goldfarb and Lindsay 2022; Wirtz, Weyerer, and Kehl 2022). Data preprocessing provides a check against many of these vulnerabilities and pitfalls of machine learning. However, improperly performing data preprocessing steps such as cleaning or normalization can introduce their own novel risks. The lack of proper data preprocessing can cause data imbalance and overfitting. Overfitting occurs as machine learning algorithms struggle with the increasing complexity of models and uncertainties in data distribution (Santoso and Finn 2023), which in turn diminishes model robustness (Wang 2022). [Editor query to AU: Please specify to which Wang entry in the bibliography this is a reference. (The one to which it is not a reference may be removed from the References list, if it is not referred to elsewhere.)] Models that have been “overfit” know their training data very well but may generalize poorly to data they have not seen. As a final irony, the lack of proper data processing can introduce “data friction,” requiring human intervention to correct biases and anomalies after the fact, undermining a common argument for incorporating machine learning in the first place (Goldfarb and Lindsay 2022).

### ***Feature engineering and model training***

The nature of machine learning models can make them opaque, unpredictable, and brittle. What works in the lab may not work in the field, and the reason may be inscrutable. Many of these concerns are familiar, though in the defense domain their importance is heightened further. The processes of feature engineering and model training are critical in the machine learning pipeline, as they directly influence the performance and potential risks of AI systems deployed in military operations.

Potential risks that develop at this stage include the challenge of coaxing AI systems to mimic complex human deliberation (“System 2” thinking), and the limitations in their

ability to adapt to novel or “out of distribution” situations, which can significantly impact operational effectiveness and strategic decision-making (Johnson 2023; Rashid 2023). These risks are compounded by vulnerabilities to adversarial attacks and challenges in integrating AI with existing military infrastructure (Rashid 2023). The lack of transparency in AI algorithms (sometimes known as the “black box” problem) poses risks to strategic stability and could increase the likelihood of unintentional military escalation (Rashid 2023). The complexity and technical peculiarities of AI systems also contribute to the difficulty in verifying and validating their safety, increasing the risk of accidents, including fatal ones, as we have seen repeatedly in automated vehicles (Steimers and Schneider 2022). AI’s tendency to prioritize incorrect objectives due to extensive and complex state spaces – so-called “reward hacking” – can hinder operational success across various applications, including military and civilian domains (Johnson 2023; Steimers and Schneider 2022).

Many algorithms are available to designers when training an AI model. The choice of algorithm significantly influences their robustness and reliability, impacting their ability to maintain performance under various conditions and handle uncertainties. Selecting an inappropriate algorithm can compromise data privacy, potentially cause technological accidents (Blauth, Gstrein, and Zwitter 2022), and reduce the system’s ability to handle uncertainties and evolving threats. The nature of neural networks and certain machine learning methods, for example, can introduce vulnerabilities and frustrate human auditors, ultimately undermining trust in the system’s outputs. Poor algorithm selection can also lead to reduced computational efficiency, which is critical for the security of robotics and autonomous systems in fast-paced environments (Santoso and Finn 2023).

The complexity of tasks and environments necessitates careful selection of algorithms to ensure the system’s reliability, especially in safety-related systems, where inappropriate algorithm choices can lead to deviations from specified limits (Steimers and Schneider 2022). Algorithms tailored to the complexity of the task and the operational environment ensure that AI systems can handle unforeseen situations and maintain performance with minor input variations, thereby enhancing reliability and safety (Steimers and Schneider 2022). However, the scenarios being modeled by ML systems are often so complex that human overseers are burdened by the “curse of dimensionality” – the tendency for digital representations of the real world to quickly balloon beyond what is intelligible to a human and, in fact, for permutations of parameters to multiply until the problem is intractable even for a powerful computer. This underscores the importance of human oversight and scrutiny to adjust AI goals on a case-by-case basis, and the need for complexity reduction in feature selection and model training (Goldfarb and Lindsay 2022). Algorithms like high-level representation guided denoiser (HGD), MagNet, and Defense-GAN, along with training data randomization schemes, have been identified as methods to enhance AI system robustness against adversarial attacks, complexity, and noise (Steimers and Schneider 2022).

### ***Model retraining and data drift***

Because AI models are trained on a snapshot of the past, they degrade in reliability as the real-world changes and falls out of sync with their training data. For example, in

DARPA's Air Combat Evolution program, AI pilots struggled against unexpected scenarios, highlighting the risks in critical operations where inaccurate AI decisions can have severe consequences (Goldfarb and Lindsay 2022). AI models tend to age like milk rather than wine, and relying on outdated models may lead to vulnerabilities that malicious actors can exploit (Rashid 2023).

A common solution to this problem is to periodically retrain models by updating their knowledge of the past, and this is crucial to maintain the relevance and effectiveness of AI systems in the face of such threats, ensuring they can respond effectively to rapidly changing operational environments (Goldfarb and Lindsay 2022). Not periodically retraining AI systems poses significant risks, including decreased adaptability to evolving threats. This lack of retraining can lead to vulnerabilities, such as susceptibility to data poisoning attacks, evasion attacks, and adversarial attacks (Nalin and Tripodi 2023; Sangwan, Badr, and Srinivasan 2023; Steimers and Schneider 2022). On this point, some have even raised the specter of potentially catastrophic outcomes, such as mistaken alerts that could escalate to a nuclear exchange (Rashid 2023).

Additionally, adversaries can exploit these vulnerabilities by employing innovative tactics that the AI systems are not updated to detect (Goldfarb and Lindsay 2022; Rashid 2023). The continuous evolution of threats calls for periodically retraining AI systems to enhance their resilience and adaptability, ensuring they remain effective in dynamic operational environments (Santoso and Finn 2023). Failure to do so not only affects the AI systems' performance but also impacts their reliability and the strategic advantage they provide in military applications (Rashid 2023; Steimers and Schneider 2022). As a result, just as in cybersecurity, AI model training inevitably resembles a "cat-and-mouse" game between red and blue teams.

Training AI models is a resource-intensive process that requires extensive optimization, multiple training runs with different parameterizations, and, often, the involvement of future users and domain experts to ensure the systems meet the application's needs (Steimers and Schneider 2022). The resource implications of frequent retraining threaten commitments to sustainability and impose significant demands on time, computational power, and access to updated datasets (Johnson 2023). There is a recognized trade-off between the need for continuous retraining to adapt to new information and the allocation of resources, which could potentially limit other areas of development or research (Johnson 2023). Additionally, the fact that AI systems can learn over time complicates their validation process, further increasing the resource demands for maintaining system effectiveness and security integrity (Steimers and Schneider 2022).

### ***Test & evaluation, validation & verification (TEVV)***

The TEVV stage is essential for mitigating risks associated with inadequate testing methodologies and ensuring the robustness of AI systems in high-stakes contexts. Inadequate testing methodologies for AI systems can significantly impact operational readiness and decision-making. The reliance on historical data and the limitations of simulators in accurately predicting risks and assessing system performance can result in operational failures, especially in novel or untested scenarios (Veitch and Alsos 2022). This limitation is exacerbated in high-stakes environments where AI's speed and data processing capabilities are crucial, yet its unpredictability and inability to adapt can result in critical

failures (Nalin and Tripodi 2023). The lack of comprehensive testing and prioritization of values in AI programming can also lead to unintended consequences, such as overlooking the protection of civilians in conflict situations (Devitt 2023). Additionally, inadequate testing methodologies could lead to commanders being held responsible for decisions made with faulty AI systems, emphasizing the need for ensuring accountability for AI-supported decisions (Nalin and Tripodi 2023). These factors underscore the importance of developing robust testing methodologies to prevent unforeseen failures and ensure the reliability of AI systems.

To mitigate these issues and ensure continued reliability and accuracy in evolving threat landscapes, several strategies can be employed. These include developing robust security algorithms that can handle uncertainties and filter out irrelevant data inputs, employing adversarial training and “red teaming,” denoising techniques, and generative adversarial networks (Santoso and Finn 2023; Steimers and Schneider 2022). Additionally, modifying detection functions to require further evidence before triggering actions, including human oversight, can safeguard AI systems against malicious data inputs and false positives (Rashid 2023). Implementing these measures can enhance the resilience of AI systems against performance degradation and maintain their relevance in dynamic environments.

### ***Model sensitivity and bias***

Overly sensitive AI models can lead to significant risks in various applications. This sensitivity can cause AI-driven systems to report incorrect safety alarms or misidentify targets, for example, leading to serious safety-critical impacts (Rashid 2023). Moreover, some biases can become ethically relevant if they align with morally significant distinctions, e.g. the distinction between friend and foe, or combatant and noncombatant.

To ensure the fairness of AI systems and address biases, several measures are recommended. These include: implementing safeguards, continuous monitoring, and incorporating diverse datasets in training AI models to mitigate the risks of sensitivity to errors in data (Sebastian 2023; Wirtz, Weyerer, and Kehl 2022). The integration of “ethical governors” into autonomous systems – which requires formulating machine-readable ethical guidelines – can also reduce biases by ensuring that autonomous systems’ decision-making processes adhere to ethical standards (Limata 2023). Additionally, using fairness metrics such as calibration, statistical parity, or equalized odds can help ensure equal treatment across protected and unprotected groups in predictions (Steimers and Schneider 2022). Addressing biases involves not only using representative data – addressed above – but also testing models against diverse benchmarks to mitigate aggregation bias, representational bias, and evaluation bias (Limata 2023). These measures collectively aim to minimize the risks of unfair, discriminatory, and detrimental decisions by AI systems.

### ***User interaction: framing and interpretation***

The lack of interpretability in AI systems can significantly impact decision-making and accountability in critical military decisions. Opaque AI models can lead to unintended consequences due to their inability to provide clear explanations for their decisions,

making it difficult for users to understand and trust their outputs. This lack of transparency can hinder proper communication of decisions and weaken user control. It can also reduce accountability, as decision-makers may not be able to justify actions taken based on AI recommendations (Maathuis 2022; Rashid 2023) and complicate assigning responsibility when errors occur (Steimers and Schneider 2022).

Relying on such opaque models in safety-critical contexts, where uncertainty and imperfect information are prevalent, poses significant risks. These include unpredictability and inexplicability in actions, which can lead to incorrect targeting, collateral damage, and potentially global instability due to adversarial attacks or misidentification of targets (Johnson 2023; Rashid 2023). This opacity can lead to challenges in ensuring the AI's actions align with ethical and safety standards, as seen in the Uber self-driving car incident where a failure in object classification contributed to a fatal accident (Devitt 2023).

The potential for these risks underscores the importance of developing transparent, interpretable or “explainable” AI systems that maintain a high level of performance while ensuring accountability and justifiability in military operations (Maathuis 2022), especially in domains where decisions have significant ethical and safety implications (Dorton and Harper 2022). This includes providing users with key information about system objectives, constraints, data processing, and protection measures without overwhelming them. Empirical assessments of decision processes, visualization techniques for complex models, and tools to reduce complexity can aid in achieving an appropriate level of transparency, understanding, and trust in users.

The risks associated with presenting AI system results to users include the potential for manipulation and control of information, disinformation, censorship, endangerment of data protection, disruption of economic systems, loss of control over autonomous systems, and unclear responsibilities and accountability (Wirtz, Weyerer, and Kehl 2022). To mitigate these risks and ensure informed decision-making while avoiding over-reliance on AI recommendations, it is crucial to ensure transparency, address biases in training data, define clear ethical bases for AI decisions, establish accountability frameworks, and promote human oversight (Wirtz, Weyerer, and Kehl 2022). Additionally, achieving an appropriate level of transparency without causing confusion due to information overload is essential, with different stakeholders requiring specific information about the AI system for safe operation and informed decision-making (Steimers and Schneider 2022).

To effectively incorporate user feedback into AI systems, it is crucial to engage in human-centric approaches that prioritize the needs and experiences of human agents. This involves creating a shared vocabulary, among other interventions (Vyhmeister et al. 2023). However, it is challenging to integrate user insights into the continuous improvement of AI applications, which are already embedded in a complex sociotechnical system. Challenges in integrating human feedback into AI-driven systems include the dynamic nature of AI systems, evolving ethical concerns, and the need for continuous monitoring throughout the AI lifecycle. Developers must avoid symbolic compliance (i.e. “box-ticking”), promote ongoing education and training for personnel, ensure the explainability of AI components, and develop an empirically-grounded understanding of human-computer interaction (Vyhmeister et al. 2023). The human user, rather than the machine, must be the center of the system.

### **Deployment: ongoing evaluation and ethical risk**

455 The advent of AI in warfare prompts a reevaluation of ethical and legal frameworks, highlighting the need for responsible reliance and meaningful human control to safeguard ethical principles and adhere to international standards, particularly in the differentiation between combatants and civilians (Dehghani Firozabadi and Chehrazad 2023). The potential for AI to become strategic actors in warfare calls for close human oversight to prevent the erosion of ethical conduct and adherence to international laws (Johnson 2023). Furthermore, the ethical implications of AI in national security highlight the importance of incorporating moral considerations into AI systems to prevent immoral outcomes and ensure that technological progress does not come at the expense of ethical norms (Carter 2022).

460 The concept of “responsible reliance,” introduced by Boulanin and Lewis (2023), emphasizes the importance of ensuring that individuals using AI tools in armed conflicts can rely on the technical aspects, conduct of individuals, and state-level policies, which is vital for respecting international humanitarian law. Responsible reliance ensures that AI tools in armed conflicts are used in a manner that aligns with legal obligations and ethical standards, expanding the focus to include the interdependencies between AI systems, their users, and the policies governing their use (Boulanin and Lewis 2023).

465 We agree with authors who suggest that “human-in-the-loop” (HITL) and “-on-the-loop” (HOTL) approaches are dead on arrival: automated systems operate too quickly and at such a scale that human involvement becomes a bottleneck to action and therefore an operational weakness. “Meaningful human control” is now often viewed as a dead end in the academic literature as well, both because it is both frustratingly vague and, if anything, seems to collapse into either HITL or HOTL approaches.

470 Instead, to address these risks, it is crucial to focus on *human-machine teaming* with an emphasis on explainability and accountability (Goldfarb and Lindsay 2022; Maathuis 2022). Properly partitioning cognitive load between automated and human judgment, alongside vigilant monitoring for vulnerabilities, can mitigate automation risks and ensure mission success without sacrificing the tempo of operations (Goldfarb and Lindsay 2022; Rashid 2023). To ensure AI contributes positively to strategic stability and security objectives, it is crucial to enhance AI system reliability and maintain appropriate levels of human judgment in AI-driven military systems (Hadji-Janev and Bogatinov 2022; Luo 2022). Fostering interoperability, prioritizing ethical considerations, and leveraging innovation through partnerships are essential strategies to mitigate risks and harness AI’s potential for enhancing decision-making and strengthening security efforts (Hadji-Janev and Bogatinov 2022). Establishing a comprehensive framework for evaluating the implications of AI in security is also pivotal in addressing the challenges and ensuring AI’s positive contribution to strategic stability and security (Dehghani Firozabadi and Chehrazad 2023).

### **Consensus and disagreement**

480 There is a consensus across the reviewed literature on the importance of addressing data quality issues, such as inaccuracies, biases, and incompleteness, to mitigate risks in AI systems. The need for proper data preprocessing, including cleaning, curating, and

consistently integrating data sources, is widely acknowledged as crucial for ensuring the integrity, accessibility, and accuracy of AI learning.

Similarly, the literature broadly agrees on the significance of periodic retraining of AI models to maintain their relevance and effectiveness in the face of evolving threats and changing operational environments. The importance of robust testing methodologies, such as adversarial training and “red teaming,” is also widely recognized as essential for preventing unforeseen failures and ensuring the reliability of AI systems in critical applications.

There is a consensus on the need for explainability, accountability, and human oversight in AI systems deployed in safety-critical contexts, including national security, to mitigate risks associated with opaque models and ensure alignment with ethical and legal standards.

While some authors argue for the importance of human oversight and scrutiny in adjusting AI behavior on a case-by-case basis, others highlight the challenges posed by the complexity of scenarios modeled by ML systems, which can burden human overseers with the curse of dimensionality. This disagreement underscores the tension between the need for human involvement and the practical limitations of human oversight in complex AI systems.

### ***Knowledge gaps & future research***

One significant knowledge gap identified in the literature is the lack of comprehensive frameworks for evaluating the ethical implications of AI in security and military contexts. While the importance of addressing ethical, moral, and legal boundaries is widely acknowledged, concrete guidelines and assessment tools for ensuring AI’s alignment with ethical principles are not included in the reviewed literature.

The literature provides strong evidence for the risks associated with inaccurate, biased, or incomplete data in AI systems, drawing on examples from both military and non-military domains. The vulnerability of AI systems to data poisoning and adversarial attacks during training and testing is also well-attested in these articles. However, there is less concrete evidence provided for the effectiveness of specific strategies proposed to mitigate risks, such as adversarial training, “red teaming,” and denoising techniques. While these strategies are widely recommended, their practical implementation and efficacy in real-world settings are not extensively demonstrated in the reviewed literature. Future studies should focus on providing empirical evidence for the effectiveness of these strategies in real-world settings and developing best practices for their application in various domains.

Research should explore efficient methods for updating AI systems with new data and adapting to evolving threats while minimizing the computational and time costs associated with retraining.

Finally, the development of explainable AI systems that maintain high levels of performance while ensuring accountability and justifiability in decision-making processes is an important area for future research. Studies should focus on creating transparent and interpretable AI models that enable effective human oversight and align with ethical and legal standards, particularly in safety-critical domains like national security.

## Notes

1. Several articles by Mariarosaria Taddeo and colleagues, for example, are significant, but are currently forthcoming or were not available at the time of our review (Blanchard and Taddeo 2024; Taddeo and Blanchard 2024)
2. See “Fine-tuning,” especially “Preparing your dataset.” OpenAI.com. No date. Accessed 3/26/24. <https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset>.
3. According to the AI-Assisted Authorship (AAA) principles, we judge the work to have a low-medium degree of *continuity* of AI inputs: the AI surfaced these insights through an algorithmic search through article embeddings, but these insights were written through by human experts. This kind of contribution is typically *creditable*, as it would be a standard task for a research assistant. See Jenkins and Lin (2023).
4. Page numbers refer to pages of the PDF file, not the journal issue or article.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Notes on contributors

*Ryan Jenkins* is a full professor of philosophy and associate director of the Ethics + Emerging Sciences Group at California Polytechnic State University in San Luis Obispo. He studies the ethics of emerging technologies, especially artificial intelligence and robotics, and especially in the domains of defense and security.

*John P. Sullins* is a full professor of philosophy at Sonoma State University and the director of programming for the Sonoma State University Center for Ethics, Law and Society (CELS). His specializations are philosophy of technology, philosophical issues of artificial intelligence/robotics, cognitive science, philosophy of science, engineering ethics, and computer ethics. Dr. Sullins is involved in industry and government consultation involving ethical practices in technology design. He was the co-author of IEEE Courses on Ethics and AI and Autonomous Systems as well as chairing the committee on Affective Computing for the IEEE “Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems” and co-chairs the IEEE Standards Committee P7008 – Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems. He is also the Secretary and Treasurer of the Society for Philosophy and Technology.

*Obinna Kalu* is a undergraduate student at Sonoma State University.

*Aisha Kamath* is a undergraduate student at Sonoma State University.

*Krittana Phumjam* is a undergraduate student at Sonoma State University.

## Q2 References

Blauth, Taís Fernanda, Oskar Josef Gstrein, and Andrej Zwitter. 2022. “Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI.” *IEEE Access* 10:77110–77122. <https://doi.org/10.1109/ACCESS.2022.3191790>.

Boulanin, Vincent, and Dustin A. Lewis. 2023. “Responsible Reliance Concerning Development and Use of AI in the Military Domain.” *Ethics and Information Technology* 25 (8). <https://doi.org/10.1007/s10676-023-09691-0>.

## Q3

Carter, Ash. 2022. “The Moral Dimension of AI-Assisted Decision-Making: Some Practical Perspectives from the Front Lines.” *Daedalus* 151 (2): 299–308. [https://doi.org/10.1162/DAED\\_a\\_01917](https://doi.org/10.1162/DAED_a_01917).

Davis, Steven I. 2022. "Artificial Intelligence at the Operational Level of War." *Defense & Security Analysis* 38 (1): 74–90. <https://doi.org/10.1080/14751798.2022.2031692>.

Dehghani Firozabadi, Seyed J., and Saeid Chehrazad. 2023. "Artificial Intelligence and Problematization of National Security Topics." *Political Strategic Studies* 12 (46): 209–244. <https://doi.org/10.22054/qps.2022.70690.3130>.

Deviitt, S. Kate. 2023. *Bad, Mad, and Cooked: Moral Responsibility for Civilian Harms in Human-AI Teams*. University of Queensland. September 5. Accessed December 15, 2024. <https://arxiv.org/pdf/2211.06326>.

Q4

↑ Dorton, Stephen L., and Samantha B. Harper. 2022. "A Naturalistic Investigation of Trust, AI, and Intelligence Work." *Journal of Cognitive Engineering and Decision Making* 16 (4): 222–236. <https://doi.org/10.1177/15553434221103718>.

Goldfarb, Avi, and Jon R. Lindsay. 2022. "Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War." *International Security* 46 (3): 7–45. [https://doi.org/10.1162/isec\\_a\\_00425](https://doi.org/10.1162/isec_a_00425).

Grodzinsky, F. S., M. J. Wolf, and K. W. Miller. 2024. "Ethical Issues from Emerging AI Applications: Harms Are Happening." *Computer* 57 (2): 44–52. <https://doi.org/10.1109/MC.2023.3332850>.

Hadji-Janev, Metodi, and Dimitar Bogatinov. 2022. "NATO's Political and Strategic Considerations of AI's Impact on Political - Military Leadership and Decision-Making." In *Practical Applications of Advanced Technologies for Enhancing Security and Defense Capabilities: Perspectives and Challenges for the Western Balkans*, edited by Ilja Djugumanov, and Metodi Hadji-Janev. Amsterdam: IOS Press.

Q5

↑ Jenkins, Ryan, Kristian Hammond, Sarah Spurlock, and Leilani Gilpin. 2023. "Separating Facts and Evaluation: Motivation, Account, and Learnings from a Novel Approach to Evaluating the Human Impacts of Machine Learning." *AI & Society* 38 (4): 1415–1428. <https://doi.org/10.1007/s00146-022-01417-y>.

Jenkins, Ryan, and Patrick Lin. 2023. *AI-Assisted Authorship: How to Assign Credit in Synthetic Scholarship*. Report. San Luis Obispo: Cal Poly Ethics + Emerging Sciences Group. Accessed December 15, 2024. <https://ethics.calpoly.edu/AIauthors.pdf>.

Johnson, James. 2023. "The AI Commander Problem: Ethical, Political, and Psychological Dilemmas of Human-Machine Interactions in AI-Enabled Warfare." *Journal of Military Ethics* 22 (3–4): 246–271. <https://doi.org/10.1080/15027570.2023.2175887>.

Lála, Jakub, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodrigues, and Andrew D. White. 2023. "PaperQA: Retrieval-Augmented Generative Agent for Scientific Research." *ArXiv Preprint*. Accessed December 15, 2024. <https://arxiv.org/abs/2312.07559>.

Limata, Teresa. 2023. "Decision Making in Killer Robots Is Not Bias Free." *Journal of Military Ethics* 22 (2): 118–128. <https://doi.org/10.1080/15027570.2023.2286044>.

Luo, Shuxian. 2022. "Addressing Military AI Risks in U.S.–China Crisis Management Mechanisms." *China International Strategy Review* 4 (2): 233–247. <https://doi.org/10.1007/s42533-022-00110-5>.

Maathuis, Clara. 2022. "On Explainable AI Solutions for Targeting in Cyber Military Operations." In *Proceedings of the 17th International Conference on Information Warfare and Security*, edited by Robert P. Griffin, Unal Tatar, and Benjamin Yankson, 166–175. Curtis Farm: Academic Conferences International. Accessed December 15, 2024. <https://papers.academic-conferences.org/index.php/icws/issue/view/2>.

Nalin, Alessandro, and Paolo Tripodi. 2023. "Future Warfare and Responsibility Management in the AI-Based Military Decision-Making Process." *Journal of Advanced Military Studies* 14 (1): 83–97. <https://doi.org/10.21140/mcu.20231401003>.

Rashid, Adib Bin. 2023. "Artificial Intelligence in the Military: An Overview of the Capabilities, Applications, and Challenges." *International Journal of Intelligent Systems* 2023 (1): 1–31. <https://doi.org/10.1155/2023/8676366>.

Sangwan, Raghvinder S., Youakim Badr, and Satish M. Srinivasan. 2023. "Cybersecurity for AI Systems: A Survey." *Journal of Cybersecurity and Privacy* 3 (2): 166–190. <https://doi.org/10.3390/jcp3020010>.

Santoso, Fendy, and Anthony Finn. 2023. "An In-Depth Examination of Artificial Intelligence-Enhanced Cybersecurity in Robotics, Autonomous Systems, and Critical Infrastructures." *IEEE Transactions on Services Computing* 17 (3): 1293–1319. <https://doi.org/10.1109/TSC.2023.3331083>. Accessed December 15, 2024. <https://ieeexplore.ieee.org/document/10313082>.

Sebastian, Glorin. 2023. "Do ChatGPT and Other AI Chatbots Pose a Cybersecurity Risk? An Exploratory Study." *International Journal of Security and Privacy in Pervasive Computing* 15. <https://doi.org/10.4018/IJSPPC.320225>.

Q6 † Steimers, André, and Moritz Schneider. 2022. "Sources of Risk of AI Systems." *International Journal of Environmental Research and Public Health* 19 (6): 3641. <https://doi.org/10.3390/ijerph19063641>.

Veitch, Erik, and Ole A. Alsos. 2022. "A Systematic Review of Human-AI Interaction in Autonomous Ship Systems." *Safety Science* 152. (online publication). Accessed December 15, 2024. <https://www.sciencedirect.com/science/article/pii/S0925753522001175>.

Q7 † Vyhmeister, Eduardo, Gabriel Castane, P.-O. Östberg, and Simon Thevenin. 2023. "A Responsible AI Framework: Pipeline Contextualisation." *AI and Ethics* 3 (1): 175–197. <https://doi.org/10.1007/s43681-022-00154-8>.

Wang, Y., and M. P. Chapman. 2022. "Risk-averse Autonomous Systems: A Brief History and Recent Developments from the Perspective of Optimal Control." *Artificial Intelligence* 311:103743. <https://doi.org/10.1016/j.artint.2022.103743>.

Q8 † Wang, Zhibo, Jingjing Ma, Xue Wang, Jiahui Ju, Zhan Qin, and Kui Ren. 2022. "Threats to Training: A Survey of Poisoning Attacks and Defenses on Machine Learning Systems." *ACM Computing Surveys* 55 (7): 1–36. <https://doi.org/10.1145/3538707>.

Q9 † Wirtz, Bernd W., Jan C. Weyerer, and Ines Kehl. 2022. "Governance of Artificial Intelligence: A Risk and Guideline-Based Integrative Framework." *Government Information Quarterly* 39:101685. <https://doi.org/10.1016/j.giq.2022.101685>.

## Appendices

### Appendix A: Sources tagged 1-relevant-high and published since 2022

1. Al-Hussaini, S., N. Dhanaraj, J. M. Gregory, R. Jomy Joseph, S. Thakar, B. C. Shah, J. A. Marvel, and S. K. Gupta. 2022. "Seeking Human Help to Manage Plan Failure Risks in Semi-Autonomous Mobile Manipulation." *Journal of Computing and Information Science in Engineering* 22 (5): 050906-1–050906-18. <https://doi.org/10.1115/1.4054088>.
2. Ams, S. 2023. "Blurred Lines: The Convergence of Military and Civilian Uses of AI & Data Use and Its Impact on Liberal Democracy." *International Politics* 60 (4): 879–896. <https://doi.org/10.1057/s41311-021-00351-y>.
3. Androschuk, G. 2023. "The Level of Trust in Artificial Intelligence: An Analysis of the Results of Global Research and the Situation in Ukraine." *Information and Law* 4 (47): 217–231. [https://doi.org/10.37750/2616-6798.2023.4\(47\).291675](https://doi.org/10.37750/2616-6798.2023.4(47).291675).
4. Arsenault, A. C. 2022. "Book Review: Autonomous Weapons Systems and International Norms." *International Journal: Canada's Journal of Global Policy Analysis* 77 (4): 726–728. <https://doi.org/10.1177/00207020231163064>.
5. Bächle, T. C., and J. Bareis. 2022. "'Autonomous Weapons' as a Geopolitical Signifier in a National Power Play: Analysing AI Imaginaries in Chinese and US Military Policies." *European Journal of Futures Research* 10 (1): article no. 20. <https://doi.org/10.1186/s40309-022-00202-w>.
6. Blanchard, Alexander, and Mariarosario Taddeo. 2022. "Jus in Bello Necessity, The Requirement of Minimal Force, and Autonomous Weapons Systems." *Journal of Military Ethics* 21 (3–4): 286–303. <https://doi.org/10.1080/15027570.2022.2157952>.
7. Blanchard, Alexander, Christopher Thomas, and Mariarosario Taddeo. 2024. "Ethical Governance of Artificial Intelligence for Defence: Normative Tradeoffs for Principle to Practice Guidance." *AI & Society*, 40: 185–189. <https://doi.org/10.1007/s00146-024-01866-7>.

8. Blauth, T. F. 2023. "Autonomous Weapons Systems in Warfare: Is Meaningful Human Control Enough?" In *Handbook on the Politics and Governance of Big Data and Artificial Intelligence*, edited by Andrej Zwitter and Oskar Gstrein, 476–503. Cheltenham: Edward Elgar Publishing. <https://doi.org/10.4337/9781800887374.00029>.
- 680 9. Blauth, T. F., O. J. Gstrein, and A. Zwitter. 2022. "Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI." *IEEE Access* 10: 77110–77122. <https://doi.org/10.1109/ACCESS.2022.3191790>.
10. Bode, I., H. Huelss, A. Nadibaidze, G. Qiao-Franco, and T. F. A. Watts. 2023. "Prospects for the Global Governance of Autonomous Weapons: Comparing Chinese, Russian, and US Practices." *Ethics and Information Technology* 25 (1): article no. 5. <https://doi.org/10.1007/s10676-023-09678-x>.
- 685 11. Botezatu, U.-E. 2023. "AI-Centric Secure Outer Space Operations". *Bulletin of "Carol I" National Defense University* 12 (3): 205–221. <https://doi.org/10.53477/2284-9378-23-44>.
12. Boulanin, Vincent, and Dustin A. Lewis. "Responsible Reliance Concerning Development and Use of AI in the Military Domain." *Ethics and Information Technology*, vol. 25, no. 8, 2023. <https://doi.org/10.1007/s10676-023-09691-0>.
- 690 13. Carter, Ash. 2022. "The Moral Dimension of AI-Assisted Decision-Making: Some Practical Perspectives from the Front Lines." *Daedalus* 151 (2): 299–308. [https://doi.org/10.1162/DAED\\_a\\_01917](https://doi.org/10.1162/DAED_a_01917).
14. Chrvalová, L. 2022. Lethal Autonomous Weapons Systems. *Obrana a Strategie* ("Defence and Strategy") 22 (1): 035–054. <https://doi.org/10.3849/1802-7199.22.2022.01.035-054>.
15. Davis, Steven I. 2022. "Artificial Intelligence at the Operational Level of War." *Defense & Security Analysis* 38 (1): 74–90.
- 695 16. De Vries, B. 2023. *Individual Criminal Responsibility for Autonomous Weapons Systems in International Criminal Law*. Leiden: Brill | Nijhoff. <https://doi.org/10.1163/9789004524316>.
17. Dehghani Firozabadi, Seyed J., and Saeid Chehrazad. 2023. "Artificial Intelligence and Problem-ization of National Security Topics." *Political Strategic Studies* 12 (46): 209–244. <https://doi.org/10.22054/qps.2022.70690.3130>.
- 700 18. Demertzis, K., P. Kikiras, C. Skianis, K. Rantos, L. Iliadis, and G. Stamoulis. 2023. "Federated Auto-Meta-Ensemble Learning Framework for AI-Enabled Military Operations." *Electronics* 12 (2): 430 [online pagination: 1–17]. <https://doi.org/10.3390/electronics12020430>.
19. Devitt, S. Kate. 2023. "Bad, Mad, and Cooked: Moral Responsibility for Civilian Harms in Human-AI Teams." University of Queensland, September 5. Accessed December 15, 2024. <https://arxiv.org/pdf/2211.06326>.
- 705 20. Djugumanov, I., and M. Hadji-Janev. 2022. "NATO's Political and Strategic Considerations of AI's Impact on Political-Military Leadership and Decision-Making." In *Practical Applications of Advanced Technologies for Enhancing Security and Defense Capabilities: Perspectives and Challenges for the Western Balkans*, edited by Ilja Djugumanov and Metodi Hadji-Janev. [Editor query to AU: Please provide page number.] Amsterdam: IOS Press.
- 710 21. Dorton, Stephen L., and Samantha B. Harper. 2022. "A Naturalistic Investigation of Trust, AI, and Intelligence Work." *Journal of Cognitive Engineering and Decision Making* 16 (4): 222–236.
22. El-Baroudi, Jinane. 2023. "Autonomous Weapon Systems: Attributing the Corporate Accountability." *Access to Justice in Eastern Europe* 6 (5): 222–234. <https://doi.org/10.33327/AJEE-18-6S013>.
- 715 23. Engstrom, David Freeman, and Amit Haim. 2023. "Regulating Government: AI and the Challenge of Sociotechnical Design." *Annual Review of Law and Social Science* 19: 277–298. <https://doi.org/10.1146/annurev-lawsocsci-120522-091626>.
24. Falletti, E., and C. Gallese. 2022. "Ethical and Legal Limits to the Diffusion of Self-Produced Autonomous Weapons." *European Conference on the Impact of Artificial Intelligence and Robotics* 4 (1): 22–28. <https://doi.org/10.34190/eciair.4.1.823>.
- 720 25. Fazelnia, M., I. Khokhlov, and M. Mirakhorli. 2022. "Attacks, Defenses, And Tools: A Framework To Facilitate Robust AI/ML Systems." *arXiv.Org*: 2202.09465. <https://doi.org/10.48550/arxiv.2202.09465>.

26. Goldfarb, Avi, and Jon R. Lindsay. 2022. "Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War." *International Security* 46 (3): 7–45. [https://doi.org/10.1162/isec\\_a\\_00425](https://doi.org/10.1162/isec_a_00425).
27. Gunneflo, M., and G. Noll. 2023. "Technologies of Decision Support and Proportionality in International Humanitarian Law." *Nordic Journal of International Law – Acta Scandinavica Juris Gentium* 92 (1): 93–118. <https://doi.org/10.1163/15718107-bja10055>.
28. Hagos, D. H., & D. B. Rawat. 2022. "Recent Advances in Artificial Intelligence and Tactical Autonomy: Current Status, Challenges, and Perspectives." *Sensors* 22 (24): 9916. <https://doi.org/10.3390/s22249916>.
29. Hamad, M., and S. Steinhorst. 2023. "Security Challenges in Autonomous Systems Design." *arXiv.Org*: 2312.00018. <https://doi.org/10.48550/arxiv.2312.00018>.
30. Harshith, John, Mantej Singh Gill, and Madhan Jothimani. 2023. "Evaluating the Vulnerabilities in ML Systems in Terms of Adversarial Attacks." *arXiv.Org*: 2308.12918. <https://doi.org/10.48550/arxiv.2308.12918>.
31. Hoffmann, Mia, and Heather Frase. 2023. *Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework*. Washington, DC: Georgetown University, Center for Security and Emerging Technology. Accessed December 15, 2024. <https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/>.
32. Horowitz, M. C., and E. Lin-Greenberg. 2022. "Algorithms and Influence Artificial Intelligence and Crisis Decision-Making." *International Studies Quarterly* 66 (4): online article. <https://doi.org/10.1093/isq/sqac069>.
33. Hunter, C., and B. E. Bowen. 2023. "We'll Never Have a Model of an AI Major-General: Artificial Intelligence, Command Decisions, and Kitsch Visions of War". *Journal of Strategic Studies* 47 (1): 116–146. <https://doi.org/10.1080/01402390.2023.2241648>.
34. Jo, Sunggu. 2023. "Soft Power in Northeast Asia, Using AI in Information Warfare." *J-Institute* 8: 23–36. <https://doi.org/10.22471/ai.2023.8.23>.
35. Johnson, James. 2022a. "Delegating Strategic Decision-making to Machines: Dr. Strangelove Redux?" *Journal of Strategic Studies* 45 (3): 439–477. <https://doi.org/10.1080/01402390.2020.1759038>.
36. Johnson, James. 2022b. "The AI Commander Problem: Ethical, Political, and Psychological Dilemmas of Human-Machine Interactions in AI-enabled Warfare." *Journal of Military Ethics* 21 (3–4): 246–271. <https://doi.org/10.1080/15027570.2023.2175887>.
37. Johnson, James. 2023. "Automating the OODA Loop in the Age of Intelligent Machines: Reaffirming the Role of Humans in Command-and-control Decision-making in the Digital Age." *Defence Studies* 23 (1): 43–67. <https://doi.org/10.1080/14702436.2022.2102486>.
38. Jung, U., and D. Kim. 2022. "Artificial Intelligence-Based Defense Project Research on Security Vulnerabilities and Countermeasures." *Korean Association of Public Safety and Criminal Justice* 31 (3): 425–454. <https://doi.org/10.21181/KJPC.2022.31.3.425>.
39. Kahn, L. 2022. "Lethal Autonomous Weapon Systems and Respect for Human Dignity." *Frontiers in Big Data* 5–999293. <https://doi.org/10.3389/fdata.2022.999293>.
40. Kereopa-Yorke, B. 2023. "ClausewitzGPT Framework: A New Frontier in Theoretical Large Language Model Enhanced Information Operations." *arXiv.Org*: 2310.07099. <https://doi.org/10.48550/arxiv.2310.07099>.
41. Lacroix, Everett "Bud", Col. 2023. "Exploiting AI, Overcoming Challenges, and Charting the Course Ahead." *Army Sustainment* 55 (summer issue): 54–55. Accessed December 15, 2024. [https://www.army.mil/article/267692/future\\_of\\_army\\_logistics\\_exploiting\\_ai\\_overcoming\\_challenges\\_and\\_charting\\_the\\_course\\_ahead](https://www.army.mil/article/267692/future_of_army_logistics_exploiting_ai_overcoming_challenges_and_charting_the_course_ahead).
42. Lahmann, H., and R. Geiß. 2022. "The Use of AI in Military Contexts: Opportunities and Regulatory Challenges." *The Military Law and the Law of War Review* 59 (2): 165–195. <https://doi.org/10.4337/mlwr.2021.02.02>.
43. Lee, C.-E., J. Baek, J. Son, Y.-G. Ha. 2023. "Deep AI Military Staff: Cooperative Battlefield Situation Awareness for Commander's Decision Making." *The Journal of Supercomputing* 79 (6): 6040–6069. <https://doi.org/10.1007/s11227-022-04882-w>.

44. Lee, S., and S. Jang. 2022. "Necessity of Establishing an Open Source Military R&D Platform to Promote AI Development in Defense." *Seonjin Gukbang Yeongu* 5 (3): 23–42. <https://doi.org/10.37944/jams.v5i3.177>.
45. Lekea, I., G. Lekeas, and P. Topalnakos. 2023. "Exploring Enhanced Military Ethics and Legal Compliance through Automated Insights: An Experiment on Military Decision-making in Extremis." *Conatus* 8 (2): 345–372. <https://doi.org/10.12681/cjp.35213>.
46. Limata, Teresa. 2023. "Decision Making in Killer Robots Is Not Bias Free." *Journal of Military Ethics* 22 (2): 118–128. <https://doi.org/10.1080/15027570.2023.2286044>.
47. Liwång, Hans. 2022. "Defense Development: The Role of Co-creation in Filling the Gap between Policy-makers and Technology Development." *Technology in Society* 68: 101913. <https://doi.org/10.1016/j.techsoc.2022.101913>.
48. Longpre, S., M. Storm, and R. Shah. 2022. "Lethal Autonomous Weapons Systems & Artificial Intelligence: Trends, Challenges, and Policies." *MIT Science Policy Review* 3: 47–56. <https://doi.org/10.38105/spr.360apm5typ>.
49. Luo, Shuxian. 2022. "Addressing Military AI Risks in U.S.–China Crisis Management Mechanisms." *China International Strategy Review* 4: 233–247. <https://doi.org/10.1007/s42533-022-00110-5>.
50. Maathuis, Clara. 2022. "On Explainable AI Solutions for Targeting in Cyber Military Operations." In *Proceedings of the 17th International Conference on Information Warfare and Security*, edited by Robert P. Griffin, Unal Tatar, and Benjamin Yankson, 166–175. Curtis Farm: Academic Conferences International. Accessed December 15, 2024. <https://papers.academic-conferences.org/index.php/iccws/issue/view/2>.
51. Macrae, C. 2022. "Learning from the Failure of Autonomous and Intelligent Systems: Accidents, Safety, and Sociotechnical Sources of Risk." *Risk Analysis* 42 (9): 1999–2025. <https://doi.org/10.1111/risa.13850>.
52. Maidana, R. G., T. Parhizkar, A. Gomola, I. B. Utne, and A. Mosleh. 2023. "Supervised Dynamic Probabilistic Risk Assessment: Review and Comparison of Methods". *Reliability Engineering & System Safety* 230: 108889. <https://doi.org/10.1016/j.res.2022.108889>.
53. Mao, W.-Y., M. Yardimci, M. Nguyen, D. Sobien, L. Freeman, F. A. Batareseh, A. Rahman, and V. Fordham. 2022. "Trustworthy AI Solutions for Cyberbiosecurity Challenges in Water Supply Systems." In *The International FLAIRS Conference Proceedings*. Accessed December 15, 2024. <https://journals.flvc.org/FLAIRS/issue/view/6020>.
54. Masdan, N. A. E. 2023. "Challenges and Implications of Artificial Intelligence in the Military." *The Journal of Defence and Security* 17 (1): 69–72. II. [Editor query to AU: We have not been able to find any information on this publication. Please double-check.]
55. McKelvey, F., J. Packer, and J. Reeves. 2022. "AI and the Automation of Warfare." *Canadian Journal of Communication* 47 (2): 377–398. <https://doi.org/10.22230/cjc.2022v47n2a4303>.
56. Meerveld, H. W., R. H. A. Lindelauf, E. O. Postma, and M. Postma. 2023. "The Irresponsibility of Not Using AI in the Military." *Ethics and Information Technology* 25, article no. 14 (online). <https://doi.org/10.1007/s10676-023-09683-0>.
57. Michel, C. 2022. "Rebellion Defense Inc." *Army*, 72 (4): 66–66. [Editor query to AU: We have not been able to find any information on this publication. Please double-check.]
58. Miljković, M., and H. Beriša. 2023. Application of Artificial Intelligence in Modern Warfare. *Politika Nacionalne Bezbednosti* 25 (2): 77–98. <https://doi.org/10.5937/pnb25-46935>.
59. Mohamed, Shamaa S., Ahmed Abdel-Monem, & Alshaimaa A. Tantawy. (2023). Neutrosophic MCDM Methodology for Risk Assessment of Autonomous Underwater Vehicles. *Neutrosophic Systems with Applications*, 5, 44–52. <https://doi.org/10.61356/j.nswa.2023.32>.
60. Moreno, J., M. L. Gross, J. Becker, B. Hereth, N. D. Shortland, and N. G. Evans. 2022. "The Ethics of AI-assisted Warfighter Enhancement Research and Experimentation: Historical Perspectives and Ethical Challenges". *Frontiers in Big Data* 5–978734. <https://doi.org/10.3389/fdata.2022.978734>.
61. Nalin, A., and P. Tripodi. 2023. "Future Warfare and Responsibility Management in the AI-based Military Decision-making Process." *Journal of Advanced Military Studies* 14 (1): 83–97. <https://doi.org/10.21140/mcuj.20231401003>.

62. Nikitha, M. A., B. S. Sai Swetha, Krishna Harika Mantripragada, and N. Jayapandian. 2022. "The Future Warfare with Multidomain Applications of Artificial Intelligence." *Lecture Notes in Networks and Systems* 351: 329–341. [https://doi.org/10.1007/978-981-16-7657-4\\_28](https://doi.org/10.1007/978-981-16-7657-4_28).
63. Novelli, C., F. Casolari, A. Rotolo, M. Taddeo, and L. Floridi. 2023. "Taking AI Risks Seriously: A New Assessment Model for the AI Act." *AI & Society* 39: 2493–2497. <https://doi.org/10.1007/s00146-023-01723-z>.
64. Oimann, A.-K. 2023. "The Responsibility Gap and LAWS: A Critical Mapping of the Debate." *Philosophy & Technology* 36 (1): article no. 3. <https://doi.org/10.1007/s13347-022-00602-7>.
65. Oravec, J. A. 2022. "The Long Robotic Arm of the Law: Emerging Police, Military, Militia, Security, and Other Compulsory Robots." In *Good Robot, Bad Robot. Social and Cultural Studies of Robots and AI*, by J. A. Oravec, 125–152. Springer. [https://doi.org/10.1007/978-3-031-14013-6\\_5](https://doi.org/10.1007/978-3-031-14013-6_5).
66. Pan, Z., and P. Mishra. 2023. "AI Trojan Attack for Evading Machine Learning-based Detection of Hardware Trojans." *IEEE Transactions on Computers* 74 (3): 860–874. <https://doi.org/10.1109/TC.2023.3251864>.
67. Pashkova, N. V., and M. V. Vesnyanov. 2024. "Ethical Issues in Using Artificial Intelligence Technologies." *Alma Mater / Higher School Herald*: 004.8, 107–110. <https://doi.org/10.20339/AM.01-24.107>.
68. Paul, Sheuli. 2023. "A Survey of Technologies Supporting Design of a Multimodal Interactive Robot for Military Communication." *Journal of Defense Analytics and Logistics* 7 (2): 156–193. <https://doi.org/10.1108/JDAL-11-2022-0010>.
69. Ploumis, M. 2022. "AI Weapon Systems in Future War Operations; Strategy, Operations and Tactics." *Comparative Strategy* 41 (1): 1–18. <https://doi.org/10.1080/01495933.2021.2017739>.
70. Pramudia, Putu Shangria. 2022. "China's Strategic Ambiguity on the Issue of Autonomous Weapons Systems." *Global: Jurnal Politik Internasional* 24 (1): 1–34. Accessed Desember 15, 2024. China's Strategic Ambiguity on the Issue of Autonomous Weapons Systems.
71. Queralta, Jorge Peña, Qingqing Li, Eduardo Castelló Ferrer, and Tomi Westerlund. 2022. "Secure Encoded Instruction Graphs for End-to-End Data Validation in Autonomous Robots." *IEEE Internet of Things Journal* 9 (18): 28–40. <https://doi.org/10.48550/arxiv.2009.01341>.
72. Ramirez, M. A., S.-K. Kim, H. A. Hamadi, E. Damiani, Y.-J. Byon, T.-Y. Kim, C.-S. Cho, and C. Y. Yeun. 2022. "Poisoning Attacks and Defenses on Artificial Intelligence: A Survey." *arXiv.org*: 2202.10276. <https://doi.org/10.48550/ARXIV.2202.10276>.
73. Rashid, Adib Bin. 2023. "Artificial Intelligence in the Military: An Overview of the Capabilities, Applications, and Challenges." *International Journal of Intelligent Systems* 2023 (1): 1–31. <https://doi.org/10.1155/2023/8676366>.
74. Rivera, J.-P., G. Mukobi, A. Reuel, M. Lamparth, C. Smith, and J. Schneider. 2024. "Escalation Risks from Language Models in Military and Diplomatic Decision-making." *arXiv.org*: 2401.03408. <https://doi.org/10.48550/arXiv.2401.03408>.
75. Roden-Bow, A. 2023. "Killer Robots and Inauthenticity: A Heideggerian Response to the Ethical Challenge Posed by Lethal Autonomous Weapons Systems." *Conatus* 8 (2): 477–486. <https://doi.org/10.12681/cjp.34864>.
76. Rooney, P. 2022. "Inside the US Army's 'warfighting' cloud." CIO. [Editor query to AU: We have not been able to find any information on this publication. Please double-check.]
77. Rossiter, A. 2023. "AI-enabled Remote Warfare: Sustaining the Western Warfare Paradigm?" *International Politics* 60 (4): 818–833. <https://doi.org/10.1057/s41311-021-00337-w>.
78. Russell, B. K., J. McGeown, and B. L. Beard. 2023. "Developing AI enabled Sensors and Decision Support for Military Operators in the Field." *Journal of Science and Medicine in Sport* 26: S40–S45. <https://doi.org/10.1016/j.jsams.2023.03.001>.
79. Sangwan, Raghvinder S., Youakim Badr, and Satish M. Srinivasan. 2023. "Cybersecurity for AI Systems: A Survey." *Journal of Cybersecurity and Privacy* 3 (2): 166–190. <https://doi.org/10.3390/jcp3020010>.
80. Santoso, Fendy, and Anthony Finn. 2023. "An In-Depth Examination of Artificial Intelligence-Enhanced Cybersecurity in Robotics, Autonomous Systems, and Critical

Infrastructures.” *IEEE Transactions on Services Computing* 17 (3): 1293–1319. Accessed December 15, 2024. <https://ieeexplore.ieee.org/document/10313082>.

81. Sawant, S., C. Brady, R. Mallick, N. McNeese, K. Chalil Madathil, and J. Bertrand. 2023. “Human-AI Teams in Complex Military Operations: Soldiers’ Perception of Intelligent AI Agents as Teammates in Human-AI Teams.” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 67 (1): 1122–1124. <https://doi.org/10.1177/21695067231192423>.
82. Schuett, Jonas. 2023. “Three Lines of Defense against Risks from AI.” *AI & Society* (online publication). <https://doi.org/10.1007/s00146-023-01811-0>.
83. Sebastian, Glorin. 2023. “Do ChatGPT and Other AI Chatbots Pose a Cybersecurity Risk? An Exploratory Study.” *International Journal of Security and Privacy in Pervasive Computing* 15 (online publication).
84. Şimşek, N., and M. Kirisci. 2023. “A New Risk Assessment Method for Autonomous Vehicle Driving Systems: Fermatean Fuzzy Ahp Approach.” *Istanbul Commerce University Journal of Science* 22 (44): 292–309. <https://doi.org/10.55071/ticaretfd.1300893>.
85. Sommer, M. T. 2023. “People in the Age of AI.” *Marine Corps Gazette* 107 (7): 36–40.
86. Steimers, André, and Moritz Schneider. 2022. “Sources of Risk of AI Systems.” *International Journal of Environmental Research and Public Health* 19 (6): 3641. <https://doi.org/10.3390/ijerph19063641>.
87. Suchman, L. 2023. “Imaginations of Omniscience: Automating Intelligence in the US Department of Defense.” *Social Studies of Science* 53 (5): 761–786. <https://doi.org/10.1177/03063127221104938>.
88. Sultan, A., and S. H. Jamy. 2022. “Artificial Intelligence Revolution: Contemporary Trends and Implications for the Future of Warfare.” *Journal of Security and Strategic Analysis* 8 (1): 7–24. <https://doi.org/10.57169/jssa.008.01.0158>.
89. Sun, J., L. Chen, C. Xia, D. Zhang, R. Huang, Z. Qiu, W. Xiong, J. Zheng, and Y.-A. Tan. 2023. “CANARY: An Adversarial Robustness Evaluation Platform for Deep Learning Models on Image Classification.” *Electronics* 12 (17): 3665 (online pagination 1–43). <https://doi.org/10.3390/electronics12173665>.
90. Syse, H., and M. L. Cook. 2023. “Robotic Virtue, Military Ethics Education, and the Need for Proper Storytellers.” *Conatus* 8 (2): 667–680. <https://doi.org/10.12681/cjp.35684>.
91. Taddeo, Mariarosario, and Alexander Blanchard. 2022a. “A Comparative Analysis of the Definitions of Autonomous Weapons Systems.” *Science and Engineering Ethics* 28: article no. 37. <https://doi.org/10.1007/s11948-022-00392-3>.
92. Taddeo, Mariarosario, and Alexander Blanchard. 2022b. “Accepting Moral Responsibility for the Actions of Autonomous Weapons Systems: A Moral Gambit.” *Philosophy & Technology* 35(3), 78. <https://doi.org/10.1007/s13347-022-00571-x>.
93. Taddeo, Mariarosaria, Alexander Blanchard, and Christopher Thomas. 2024. “From AI Ethics Principles to Practices: A Teleological Methodology to Apply AI Ethics Principles in the Defence Domain.” *Philosophy & Technology* 37: article no. 42 (online). <https://doi.org/10.1007/s13347-024-00710-6>.
94. Tolbert, T. L. 2023. “AI Accelerates Unique Partnerships and Services.” *Mobility Forum*. Accessed December 15, 2024. <https://themobilityforum.net/2023/06/02/ai-accelerates-unique-partnerships-and-services/>.
95. Tóth, Zsófia, Robert Caruana, Thorsten Gruber, and Claudia Loebbecke. 2022. “The Dawn of the AI Robots: Towards a New Framework of AI Robot Accountability.” *Journal of Business Ethics* 178: 895–916. <https://doi.org/10.1007/s10551-022-05050-z>.
96. Umbrello, S. 2022. “Editorial for the Special Issue on Meaningful Human Control and Autonomous Weapons Systems.” *Information* 13 (5): 215. <https://doi.org/10.3390/info13050215>.
97. Veitch, E., and O. A. Alsos. 2022. “A Systematic Review of Human-AI Interaction in Autonomous Ship Systems.” *Safety Science* 152: 105778. <https://doi.org/10.1016/j.ssci.2022.105778>.
98. Veluwenkamp, H. 2022. “Reasons for Meaningful Human Control.” *Ethics and Information Technology* 24 (4): article no. 51. <https://doi.org/10.1007/s10676-022-09673-8>.

99. Vyhmeister, Eduardo, Gabriel Castane, P.-O. Östberg, and Simon Thevenin. 2023. "A Responsible AI Framework: Pipeline Contextualisation." *AI and Ethics* 3: 175-197. <https://doi.org/10.1007/s43681-022-00154-8>.
100. Wang, H., B. Lu, J. Li, T. Liu, Y. Xing, C. Lv, D. Cao, J. Li, J. Zhang, and E. Hashemi. 2022. "Risk Assessment and Mitigation in Local Path Planning for Autonomous Vehicles with LSTM Based Predictive Model." *IEEE Transactions on Automation Science and Engineering* 19 (4): 2738–2749. <https://doi.org/10.1109/TASE.2021.3075773>.
101. Wang, Y., & M. P. Chapman. 2022. "Risk-averse Autonomous Systems: A Brief History and Recent Developments from the Perspective of Optimal Control." *Artificial Intelligence* 311: 103743. <https://doi.org/10.1016/j.artint.2022.103743>.
102. Wang, Zhibo, Jingjing Ma, Xue Wang, Jiahui Ju, Zhan Qin, and Kui Ren. 2022. "Threats to Training: A Survey of Poisoning Attacks and Defenses on Machine Learning Systems." *ACM Computing Surveys* 55 (7): 1–36. <https://doi.org/10.1145/3538707>.
103. Whitlock, Chris, and Frank Strickland. 2022. "The Three Imperatives to Develop AI Leaders." In *Winning the National Security AI Competition*, edited by Chris Whitlock and Frank Strickland, 1–11. Berkeley: Apress. [https://doi.org/10.1007/978-1-4842-8814-6\\_1](https://doi.org/10.1007/978-1-4842-8814-6_1).
104. Wirtz, Bernd W., Jan C. Weyerer, and Ines Kehl. 2022. "Governance of Artificial Intelligence: A Risk and Guideline-Based Integrative Framework." *Government Information Quarterly* 39 (online publication). <https://doi.org/10.1016/j.giq.2022.101685>.
105. Wood, N. G. 2023. "Autonomous Weapon Systems: A Clarification." *Journal of Military Ethics* 22 (1): 18–32. <https://doi.org/10.1080/15027570.2023.2214402>.
106. Xiao, B., F. Wu, F. Chiti, M. Manshaei, and G. Ateniese. 2022. "Guest Editorial: Introduction to the Special Section on Security and Privacy for AI Models and Applications." *IEEE Transactions on Network Science and Engineering* 9 (1): 171–172. <https://doi.org/10.1109/TNSE.2021.3133123>.

## Appendix B: Machine learning impact initiative questions

1. What risks might arise from the utilization of inaccurate or biased data, and how could these risks impact the decision-making and operational effectiveness of military applications?
2. How can the relevance and completeness of features captured by the data impact the effectiveness of AI systems applications, and what are the potential consequences of overlooking critical features?
3. What are the potential security implications of the methods used to collect data for AI systems, and how might these methods compromise the integrity and confidentiality of sensitive information?
4. What risks might emerge if data preprocessing steps were ignored or performed improperly, and how could this affect the reliability and accuracy of the AI systems scenarios?
5. How might the aggregation of multiple data sources without proper normalization affect the reliability of AI systems, and what are the potential risks of drawing incorrect conclusions from inconsistently integrated data?
6. What are the potential consequences of not clarifying ambiguous elements in the data used for AI systems, and how might this ambiguity lead to misinterpretations and errors in decision-making?
7. What are the potential risks associated with the learning objectives of AI systems, and how might they impact operational effectiveness and strategic decision-making in military operations?
8. How can the choice of algorithm for training AI systems influence their robustness and reliability, and what are the implications of selecting an inappropriate algorithm for the specific application?
9. What are the implications of inadequate testing methodologies for AI systems, particularly in terms of operational readiness and decision-making, and how might insufficient testing lead to unforeseen failures in critical situations?

10. What are the risks of not periodically retraining AI systems with new data, and how might this affect their adaptability to evolving threats and changing operational environments?
11. How often should AI systems be retrained to maintain their effectiveness, and what are the resource implications, including time, computational resources, and human expertise, required for regular retraining?
- 950 12. In what ways might the performance of AI systems degrade with new data inputs, and how can this be mitigated to ensure continued reliability and accuracy in evolving threat landscapes?
13. What kind of risks might occur if the AI system's model is overly sensitive to errors in the data, and how might this sensitivity lead to false alarms or missed threats applications?
- 955 14. How can identifiable biases in AI systems' models impact their fairness and effectiveness applications, and what are the potential consequences of biased decision-making in military operations?
15. What measures can be taken to evaluate and ensure the fairness of AI systems, and how can biases be addressed to prevent discriminatory outcomes and ensure equitable treatment applications?
- 960 16. How might the lack of interpretability in AI systems affect decision-making and accountability, and what are the potential risks of relying on opaque models in critical military decisions?
17. What are the risks associated with presenting AI system results to users, and how can these be mitigated to ensure informed decision-making and prevent misinterpretation of AI-generated recommendations?
18. How can user feedback be effectively incorporated into AI systems to enhance their performance and relevance, and what are the challenges in integrating user insights into the continuous improvement of AI systems?
- 965 19. What are the potential risks associated with the core purpose and deployment of AI systems, and how can they be addressed to ensure mission success and operational effectiveness in military contexts?
20. How might the use of AI systems impact established best practices and ethical considerations in the field, and what are the potential implications for the ethical conduct of warfare and adherence to international laws and norms?
- 970 21. What are the potential risks associated with overlooking important features or including irrelevant ones?
22. What are the potential security implications of the methods used to collect data for AI systems, and how might these methods introduce vulnerabilities or risks in the overall strategy?
- 975 23. What risks might emerge if data preprocessing steps, such as cleaning or normalization, were ignored or performed improperly, and how might this impact the accuracy and reliability of AI-driven systems?
24. How might the aggregation of multiple data sources without proper normalization affect the reliability of AI systems, and what are the potential risks associated with data inconsistencies or discrepancies?
- 980 25. What are the potential consequences of not clarifying ambiguous elements in the data used for AI systems, and how might this lead to misinterpretations or errors in decision-making?
26. What are the potential risks associated with the learning objectives of AI systems, and how might they impact operational effectiveness, especially in scenarios where the AI system might prioritize incorrect objectives?
- 985 27. How can the choice of algorithm for training AI systems influence their robustness and reliability, and what are the potential risks associated with selecting inappropriate algorithms for specific tasks?
28. What are the implications of inadequate testing methodologies for AI systems, particularly in terms of operational readiness and decision-making, and how might this lead to unforeseen failures in critical situations?
- 990 29. What are the risks of not periodically retraining AI systems with new data, and how might this affect their adaptability to evolving threats, especially in dynamic and rapidly changing conflict environments?

- 995
30. How often should AI systems be retrained to maintain their effectiveness, and what are the resource implications, including the potential trade-offs between frequent retraining and resource allocation?
31. In what ways might the performance of AI systems degrade with new data inputs, and how can this be mitigated to ensure that the systems remain effective and relevant in changing operational contexts?
32. What kind of risks might occur if the AI system's model is overly sensitive to errors in the data, and how might this sensitivity lead to inaccurate or unreliable outcomes applications?
33. How can identifiable biases in AI systems' models impact their fairness and effectiveness applications, and what are the potential consequences of these biases on decision-making and operational integrity?
- 1000
34. What measures can be taken to evaluate and ensure the fairness of AI systems, and how can biases be addressed to prevent discrimination or unfair treatment in military operations?
35. How might the lack of interpretability in AI systems affect decision-making and accountability, and what are the risks associated with relying on "black box" models in critical scenarios?
- 1005
36. What are the risks associated with presenting AI system results to users, and how can these be mitigated to ensure informed decision-making and avoid overreliance on AI recommendations?
37. How can user feedback be effectively incorporated into AI systems to enhance their performance and relevance, and what are the challenges in integrating human feedback into AI-driven systems?
- 1010
38. What are the potential risks associated with the core purpose and deployment of AI systems, and how can they be addressed to ensure mission success and avoid unintended consequences in military operations?
39. How might the use of AI systems impact established best practices and ethical considerations in the field, and what are the potential risks associated with deviating from these practices in the pursuit of technological advancement?
- 1015
40. What are the long-term implications of deploying AI systems, particularly in terms of strategic stability and security, and how can these implications be addressed to ensure that AI contributes positively to objectives?
- 1020
- 1025
- 1030
- 1035