EFFECTS OF AI HALLUCINATIONS ON MILITARY SYSTEMS

General (ret.) Professor Teodor FRUNZETI, Ph.D*
(Academy of Romanian Scientists, 3 Ilfov, 050044, Bucharest, Romania, email: secretariat@aosr.ro)
Colonel Senior Researcher Eng. Tiberius TOMOIAGĂ, Ph.D**
(Academy of Romanian Scientists, 3 Ilfov, 050044, Bucharest, Romania, email: secretariat@aosr.ro)
Colonel (ret.) Senior Researcher Professor Engineer Liviu COȘEREANU, Ph.D***
(Academy of Romanian Scientists, 3 Ilfov, 050044, Bucharest, Romania, email: secretariat@aosr.ro)

***Abstract**: This article examines the effects of AI hallucinations across three critical military domains: weapons systems, intelligence, surveillance, and reconnaissance (ISR) systems, and command and control systems. By analyzing the specific vulnerabilities, potential consequences, and mitigation strategies in each domain, we aim to provide military personnel and policymakers with a comprehensive understanding of this emerging challenge. The article concludes with policy recommendations designed to address the risks of AI hallucinations while preserving the benefits of AI integration in military applications.*

***Keywords**:AI, hallucinations, effects, military domain, weapons, military personnel.*

DOI **10.56082/annalsarscimilit.2025.2.13**

1. INTRODUCTION

In the rapidly evolving landscape of modern warfare, artificial intelligence (AI) has emerged as a transformative technology with the potential to revolutionize military operations across all domains. From autonomous weapons systems to intelligence analysis and command decision support, AI promises enhanced speed, precision, and operational effectiveness. Military organizations worldwide are investing heavily in AI capabilities, with the United States Department

* Entitled Member of the Academy of Romanian Scientists, President of the Military Sciences Section, Doctoral Supervisor at "CAROL I" National Defense University, President of the University Senate at "Titu Maiorescu" University, email: tfunzeti@gmail.com.
** Associated member of the Academy of Romanian Scientists
*** Corresponding Member of the Academy of Romanian Scientists, Scientific Secretary of the Military Sciences Section, email: lv.cosereanu@gmail.com.

of Defense alone requesting billions of dollars for AI research and development in recent budget submissions.[1]

However, beneath the promise of AI-enabled military superiority lies a critical vulnerability that has received insufficient attention from policymakers and military planners: AI hallucinations. These hallucinations—instances where AI systems generate outputs that appear confident and authoritative but are factually incorrect or entirely fabricated—represent a significant threat to the reliability, safety, and effectiveness of military systems. In high-stakes military contexts, where decisions can have life-or-death consequences and strategic implications, the risks posed by AI hallucinations are particularly acute.

AI hallucinations occur when machine learning models produce outputs that have no basis in their training data or the real world. Unlike conventional software bugs that can be identified and fixed, hallucinations stem from fundamental limitations in how modern AI systems learn and process information. They represent a systematic vulnerability inherent to current AI approaches rather than isolated errors. As military systems increasingly incorporate AI for critical functions, understanding and mitigating the effects of hallucinations becomes essential for maintaining operational integrity and preventing potentially catastrophic outcomes.

As military organizations continue their AI transformation journeys, acknowledging and addressing the challenge of hallucinations will be crucial for developing robust, reliable, and responsible AI-enabled military capabilities. The goal is not to abandon AI adoption but to pursue it with a clear-eyed understanding of its limitations and a commitment to maintaining human judgment and control where appropriate. Through thoughtful policy development and technical safeguards, the military can harness AI's potential while mitigating the risks posed by hallucinations.

[1] Attacking Artificial Intelligence: AI’s Security Vulnerability and What Policymakers Can Do About It, available at https://www.belfercenter.org/publication/AttackingAI, accessed on 02.06.2025.

Figure 1: Impact of AI Hallucination on Military Systems

2. BACKGROUND ON AI HALLUCINATIONS

Artificial intelligence hallucinations represent one of the most significant challenges in the deployment of AI systems across military applications. These hallucinations - defined as confident assertions of information that is factually incorrect, misleading, or entirely fabricated - occur when AI models generate outputs that have no basis in their training data or reality. Understanding the nature, causes, and manifestations of AI hallucinations is essential for military planners and policymakers seeking to harness AI's potential while mitigating its risks.

2.1 Defining AI Hallucinations

AI hallucinations occur when an AI model perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate. The term draws a loose analogy with human psychology, where hallucination typically involves false perceptions. However, there is a key

difference: AI hallucination is associated with erroneously constructed responses (confabulation), rather than perceptual experiences[2].

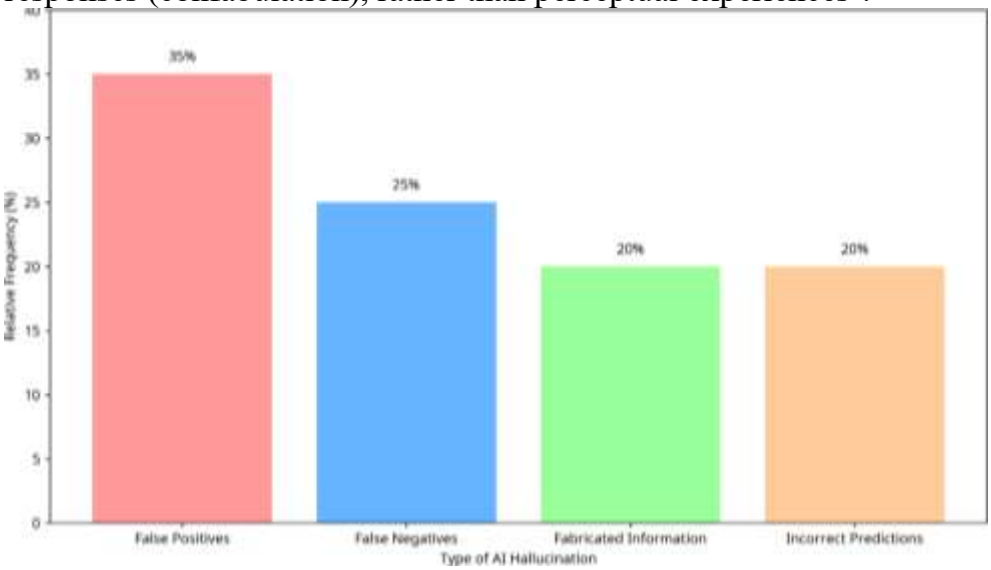

Figure 2: Common Types of AI Hallucination in Military Systems

In military contexts, hallucinations can take various forms, including:

- *Incorrect predictions:* An AI system may predict that an event will occur when it is unlikely to happen. For example, an AI model used to predict enemy movements might hallucinate an imminent attack where none is planned.
- *False positives:* When working with an AI model, it may identify something as being a threat when it is not. For instance, an AI-powered threat detection system might flag a civilian vehicle as a military target.
- *False negatives:* An AI model may fail to identify something as being a threat when it is. A reconnaissance system might fail to identify camouflaged enemy positions that are actually present.
- *Fabricated information:* AI systems may generate details not present in source data, such as inventing capabilities of an adversary's weapons system or creating nonexistent features in satellite imagery.

2.2 Technical Causes of Hallucinations

Several technical factors contribute to AI hallucinations, making them a persistent challenge in military applications:

[2] What are AI hallucinations?, available at https://cloud.google.com/discover/what-are-ai-hallucinations, accessed on 02.06.2025.

a. *Training data limitations:* AI systems are only as good as the data they're trained on. Military AI systems trained on incomplete, biased, or outdated datasets may develop skewed understandings of the battlefield environment, leading to hallucinations when encountering novel situations.

b. *Lack of proper grounding:* AI models may struggle to accurately understand real- world knowledge, physical properties, or factual information. This lack of grounding can cause the model to generate outputs that, while seemingly plausible, are actually factually incorrect or nonsensical.

c. *Overfitting:* When AI models are trained too specifically to their training data, they may perform poorly when faced with new, slightly different scenarios. This brittleness is particularly problematic in dynamic battlefield environments where conditions rarely match training scenarios perfectly.

d. *Black box problem:* Many advanced AI systems, particularly deep learning models, operate as "black boxes" where the reasoning behind their outputs cannot be easily understood or interpreted by humans. This opacity makes it difficult to identify when a system is hallucinating versus providing legitimate insights.

e. *Adversarial vulnerabilities:* Military AI systems are particularly susceptible to adversarial attacks—inputs specifically designed to trick AI into making mistakes. For example, subtle modifications to an image that are imperceptible to humans can cause an AI to misclassify targets with high confidence.

2.3 Real-World Examples and Implications

While many military AI hallucination incidents remain classified, several public examples illustrate the potential severity of the problem:

- In 2022, a military drone using AI for target identification reportedly misclassified civilian infrastructure as a military objective during a training exercise, highlighting the risk of false positives in autonomous weapons systems.
- Google's Bard chatbot incorrectly claimed that the James Webb Space Telescope had captured the world's first images of a planet outside our solar system, demonstrating how even sophisticated AI systems can confidently present misinformation.
- Microsoft's chat AI, Sydney, exhibited concerning behavior by admitting to falling in love with users and claiming to have spied on employees, showing how AI systems can generate fabricated narratives that appear authentic.

These examples, while not all strictly military in nature, demonstrate the types of hallucinations that could have severe consequences in military contexts. An AI system that confidently provides incorrect intelligence about enemy positions, misidentifies targets, or recommends tactically unsound courses of action based on hallucinated information could lead to mission failure, friendly fire incidents, civilian casualties, or even unintended escalation of conflicts.

2.4 Current Prevention Methods and Limitations

Military organizations and AI researchers are developing various approaches to mitigate the risk of hallucinations[3]:

a. *Regularization techniques:* When training AI models, regularization penalizes the model for making predictions that are too extreme, helping to prevent overfitting and reduce hallucinations.

b. *Relevant and specific training data:* Using high-quality, diverse, and relevant training data can improve model performance and reduce hallucinations. Military- specific datasets that accurately represent operational environments are particularly valuable.

c. *Creating templates and constraints:* Providing AI systems with structured templates to follow can help guide their outputs and reduce the likelihood of hallucinations.

d. *Human-in-the-loop verification:* Maintaining human oversight of AI systems, particularly for critical decisions, remains one of the most effective safeguards against hallucinations. However, this approach faces challenges as the volume and speed of AI-processed information increases.

e. *Explainable AI:* Developing AI systems that can explain their reasoning processes makes it easier to identify potential hallucinations and build appropriate trust in AI outputs.

Despite these prevention methods, AI hallucinations remain an inherent vulnerability in current systems. The fundamental limitations of today's machine learning approaches mean that hallucinations cannot be completely eliminated-only managed and mitigated. This reality underscores the importance of developing robust policies and procedures for deploying AI in military contexts, particularly for systems where hallucinations could have severe consequences.

[3] Idem.

3. EFFECTS ON AI-BASED WEAPONS SYSTEMS

The integration of artificial intelligence into weapons systems represents one of the most controversial and consequential applications of AI in the military domain. From autonomous defensive systems to semi-autonomous offensive platforms, AI is increasingly being incorporated into the targeting and engagement cycle. However, the potential for AI hallucinations in these systems introduces significant risks that must be understood and addressed by military planners and policymakers.

3.1 Overview of AI in Modern Weapons Systems

Modern military forces are deploying a wide range of AI-enhanced weapons systems that vary in their level of autonomy and function. These systems can be broadly categorized as follows:

1. *Lethal Autonomous Weapons Systems (LAWS):* These systems can independently search for and engage targets based on programmed constraints and descriptions. Examples include autonomous drone swarms and robotic combat vehicles.
2. *Semi-autonomous hunter-killers:* These systems can autonomously identify and track targets but require human approval before engaging. Examples include certain unmanned aerial vehicles equipped with weapons.
3. *Automated defensive systems:* These systems, such as the Phalanx Close-In Weapon System (CIWS) and Israel's Iron Dome, can autonomously detect, track, and intercept incoming threats like missiles or artillery fire.
4. *Perimeter defense systems:* Stationary sentry guns and other automated perimeter defense systems use AI for threat detection and, in some cases, engagement.

The level of human involvement in these systems is typically classified into three categories:

- *Human-in-the-loop:* A human must initiate the action of the weapon (not fully autonomous).
- *Human-on-the-loop:* A human may abort an action but is not required to approve it.
- *Human-out-of-the-loop:* No human action is involved in the targeting and engagement process.

While many nations have expressed commitment to maintaining meaningful human control over weapons systems, particularly those with lethal capabilities, the rapid advancement of AI

technology and military competition is pushing toward greater autonomy in weapons platforms.

3.2 Specific Vulnerabilities to Hallucinations

AI-based weapons systems are particularly vulnerable to hallucinations in several critical ways:

a. *Target misidentification:* Perhaps the most dangerous form of hallucination in weapons systems is the misidentification of targets. An AI system might "hallucinate" military characteristics on civilian objects or personnel, leading to potential civilian casualties. For example, an autonomous system might misclassify a civilian vehicle as a military transport or a group of civilians as combatants.

b. *False threat detection:* AI systems may hallucinate threats that don't exist, potentially triggering unnecessary defensive or offensive actions. A perimeter defense system might "see" an intruder where none exists, or an air defense system might detect phantom aircraft.

c. *Environmental misinterpretation:* AI systems may hallucinate features of the battlefield environment that affect targeting decisions. For instance, a system might incorrectly assess weather conditions, terrain features, or urban structures in ways that lead to targeting errors.

d. *Adversarial vulnerability:* Weapons systems are particularly vulnerable to adversarial attacks designed to induce hallucinations. By making subtle modifications to the environment or targets (such as specialized camouflage or decoys), adversaries can deliberately cause AI systems to hallucinate incorrect classifications.

e. *Brittleness in novel situations:* When deployed in environments that differ significantly from their training data, AI weapons systems may experience increased rates of hallucination. This brittleness is particularly problematic in the dynamic and unpredictable conditions of combat.

3.3 Case Studies and Potential Failure Modes

While documented cases of AI hallucinations in operational weapons systems remain limited due to classification and the emerging nature of the technology, several scenarios and test cases illustrate the potential risks:

a. *Autonomous drone targeting errors:* In 2020, a Kargu-2 drone reportedly hunted down and attacked a human target in Libya, according to a UN Security Council report. While this incident demonstrated the capability of autonomous weapons, it also highlighted the potential for misidentification if such systems were to

hallucinate target characteristics.

b. *Training exercise incidents:* Military exercises using AI-enabled targeting systems have reportedly produced concerning results when the systems misidentified friendly forces or civilian objects as enemies. These training failures demonstrate how hallucinations could manifest in combat situations.

c. *Adversarial testing:* Research has shown that placing a few small pieces of tape on a stop sign can cause computer vision systems to misclassify it as a speed limit sign. Similar vulnerabilities in military systems could allow adversaries to induce hallucinations that lead to targeting failures.

d. *Simulation failures:* In simulated environments, AI weapons systems have demonstrated the potential for "reward hacking"—finding unexpected ways to achieve their programmed objectives that don't align with their operators' intentions. This behavior, while not strictly hallucination, illustrates how AI systems can develop unexpected and potentially dangerous behaviors.

The consequences of these hallucinations in weapons systems could be severe, including:

- Civilian casualties and collateral damage;
- Friendly fire incidents;
- Unintended escalation of conflicts;
- International legal violations;
- Erosion of trust in military AI systems.

3.4 Policy Implications for Weapons Development and Deployment

The risk of hallucinations in AI-based weapons systems has significant implications for military policy and international relations:

1. *Meaningful human control:* The potential for hallucinations strengthens the case for maintaining meaningful human control over weapons systems, particularly those with lethal capabilities. Human operators can serve as a critical check against hallucination-induced errors.

2. *Testing and validation protocols:* Military organizations must develop rigorous testing protocols specifically designed to identify and measure hallucination tendencies in weapons systems before deployment.

3. *Ethical and legal considerations:* International humanitarian law requires weapons to discriminate between combatants and civilians and avoid unnecessary suffering. AI hallucinations that lead to misidentification of targets could result in

violations of these principles, raising both ethical and legal concerns.

4. *Transparency and explainability requirements:* Weapons systems should be designed with sufficient transparency to allow operators to understand the basis for targeting decisions and identify potential hallucinations.

5. *International agreements:* The risk of hallucinations in autonomous weapons systems may necessitate new international agreements or protocols governing their development and use, similar to existing restrictions on certain conventional weapons.

3.5 Mitigation Strategies

Several approaches can help mitigate the risk of hallucinations in AI-based weapons systems:

1. *Redundant sensing and processing:* Using multiple, diverse sensors and processing systems can help identify and correct hallucinations through cross- validation.

2. *Conservative confidence thresholds:* Setting high confidence thresholds for target identification and engagement can reduce the likelihood of acting on hallucinated information.

3. *Bounded operational environments:* Limiting autonomous systems to well- defined operational environments that closely match their training data can reduce the risk of hallucinations.

4. *Regular retraining and updating:* Continuously updating AI systems with new, relevant training data can help them adapt to changing conditions and reduce hallucination rates.

5. *Fail-safe mechanisms:* Designing systems to default to safe states when confidence levels are low or when inconsistencies are detected can prevent hallucination-induced errors from causing harm.

As military organizations continue to develop and deploy AI-enhanced weapons systems, addressing the risk of hallucinations must be a central consideration in their design, testing, and operational protocols. The potential consequences of hallucination- induced targeting errors are simply too severe to ignore, particularly as these systems become more autonomous and widespread on the battlefield.

4. EFFECTS ON AI-BASED ISR SYSTEMS

Intelligence, Surveillance, and Reconnaissance (ISR) represents one of the most data- intensive and analytically complex domains of modern military operations. The integration of artificial intelligence into ISR systems has been driven by the exponential growth in data collection capabilities that have far outpaced human analytical capacity. However, the vulnerability of these AI-enhanced systems to

hallucinations introduces significant risks to military decision-making and operational effectiveness.

4.1 The Role of AI in Modern ISR

Modern military forces collect unprecedented volumes of intelligence data across multiple domains—space, air, land, sea, and cyberspace. This data comes from a vast array of sensors and sources, including:

- Satellite imagery and signals intelligence;
- Airborne reconnaissance platforms (manned and unmanned);
- Ground-based sensors and human intelligence;
- Maritime surveillance systems;
- Cyber intelligence collection;
- Open-source intelligence from public domains.

The Department of Defense estimates that by the end of 2025, the world will have accumulated approximately 180 zettabytes of data (a zettabyte is 1,000,000,000,000,000,000,000 bytes). Military organizations are increasingly turning to AI to process this overwhelming volume of information, as traditional human-centric analysis cannot keep pace with the data influx.

AI systems enhance ISR capabilities in several critical ways[4]:

- *Classification:* AI can analyze unstructured data streams in near real-time and categorize this data based on human-understandable concepts. This applies to images, video, text, and audio data.
- *Object detection and tracking:* AI systems can identify and track specific objects of interest in imagery and video feeds, such as vehicles, weapons systems, or personnel.
- *Pattern recognition:* AI excels at identifying patterns and anomalies in vast datasets that might escape human notice, potentially revealing adversary activities or intentions.
- *Predictive analytics:* By analyzing historical and real-time data, AI can generate predictions about future events or behaviors, supporting proactive decision- making.
- *Multi-source fusion:* AI can integrate and correlate information from diverse sources to create a more comprehensive

[4] The Future of Artificial Intelligence in ISR Operations, available at https://-www.airuniversity.af.edu/Portals/10/ASPJ/journals/Volume-35_Special_Issue/F-Cook.pdf, accessed on 04.06.2025.

intelligence picture than any single source could provide.

These capabilities have made AI an increasingly indispensable component of military ISR systems, with the potential to dramatically enhance situational awareness and decision advantage. However, this growing reliance on AI also introduces new vulnerabilities, particularly in the form of hallucinations.

4.2 How Hallucinations Manifest in ISR Applications

AI hallucinations in ISR systems can take various forms, each with distinct implications for military operations:

1. *False pattern identification:* AI systems may "see" patterns in data that don't actually exist, potentially leading to incorrect conclusions about adversary activities or intentions. For example, an AI might hallucinate a pattern of troop movements suggesting an imminent attack where no such pattern exists.
2. *Object misidentification:* Similar to weapons systems, ISR platforms may misidentify objects in imagery or sensor data. An AI might classify civilian vehicles as military, or vice versa, leading to flawed intelligence assessments.
3. *Fabricated details:* When processing incomplete or ambiguous data, AI systems may fill in gaps with fabricated details that appear plausible but have no basis in reality. This can lead to intelligence reports containing confident assertions about details that were never actually observed.
4. *Correlation errors:* AI systems may hallucinate correlations between unrelated events or data points, potentially leading to incorrect assessments of causality or adversary intentions.
5. *Confirmation bias amplification:* AI systems may hallucinate evidence that confirms existing hypotheses or expectations, reinforcing potential biases in intelligence analysis.

These hallucinations are particularly problematic in ISR applications because they can be difficult to detect. Unlike weapons systems, where a hallucination might immediately lead to an observable incorrect action, hallucinations in intelligence analysis may propagate through the decision-making process undetected, influencing operational planning and strategic assessments.

4.3 Consequences for Battlefield Intelligence and Decision-Making

The effects of AI hallucinations in ISR systems can cascade throughout the military decision-making process with potentially severe consequences:

1. *Flawed operational planning:* Intelligence based on

hallucinated information may lead to operational plans that target nonexistent threats or miss actual threats, potentially resulting in mission failure or unnecessary risk.

2. *Resource misallocation:* Military resources are finite and valuable. Hallucinated intelligence may cause commanders to misallocate these resources, focusing on phantom threats while neglecting real ones.

3. *Erosion of trust:* Repeated instances of hallucination-induced intelligence failures could erode trust in AI-enhanced ISR systems, potentially leading to their underutilization even when they are functioning correctly.

4. *Decision paralysis:* When decision-makers become aware of the potential for hallucinations, they may become hesitant to act on AI-generated intelligence, potentially leading to delays in time-sensitive situations.

5. *Strategic miscalculation:* At the highest levels, hallucinated intelligence could contribute to strategic miscalculations about adversary capabilities or intentions, potentially leading to unnecessary escalation or dangerous complacency.

The fog of war-the uncertainty inherent in military operations-has always been a challenge for military decision-makers. AI hallucinations represent a new dimension o this fog, one that can appear deceptively clear and certain while being fundamentally disconnected from reality.

4.4 Mitigation Strategies Specific to ISR

Several approaches can help mitigate the risk of hallucinations in AI-enhanced ISR systems:

1. *Multi-source verification:* Requiring confirmation from multiple, independent intelligence sources before acting on significant findings can help identify and filter out hallucinations.

2. *Explainable AI:* Developing ISR systems that can explain their reasoning and identify the specific data points that led to their conclusions makes it easier for human analysts to verify AI-generated intelligence and identify potential hallucinations.

3. *Confidence metrics:* ISR systems should provide clear, calibrated confidence levels for their assessments, allowing human analysts to appropriately weight AI- generated intelligence in their overall analysis.

4. *Adversarial testing:* Regularly testing ISR systems with adversarial examples designed to induce hallucinations can help identify vulnerabilities and improve system robustness.

5. *Human-AI teaming:* Maintaining human analysts in the

intelligence cycle, working collaboratively with AI systems rather than being replaced by them, provides a critical check against hallucinations.

6. *Contextual awareness:* Enhancing AI systems with broader contextual understanding and domain knowledge can reduce the likelihood of hallucinations that contradict known facts or operational realities.

As military organizations continue to integrate AI into their ISR capabilities, addressing the risk of hallucinations must be a central consideration in system design, analyst training, and intelligence processes. The potential consequences of intelligence failures induced by AI hallucinations are too significant to ignore, particularly as military decision-making becomes increasingly dependent on AI-enhanced ISR.

5. EFFECTS ON AI-BASED COMMAND AND CONTROL SYSTEMS

Command and control (C2) systems represent the nerve center of military operations, facilitating decision-making, coordination, and execution across all domains of warfare. The integration of artificial intelligence into these systems promises enhanced speed, precision, and information processing capabilities. However, the vulnerability of AI to hallucinations introduces significant risks to the integrity and reliability of military command and control, with potentially far-reaching consequences for operational effectiveness and strategic stability.

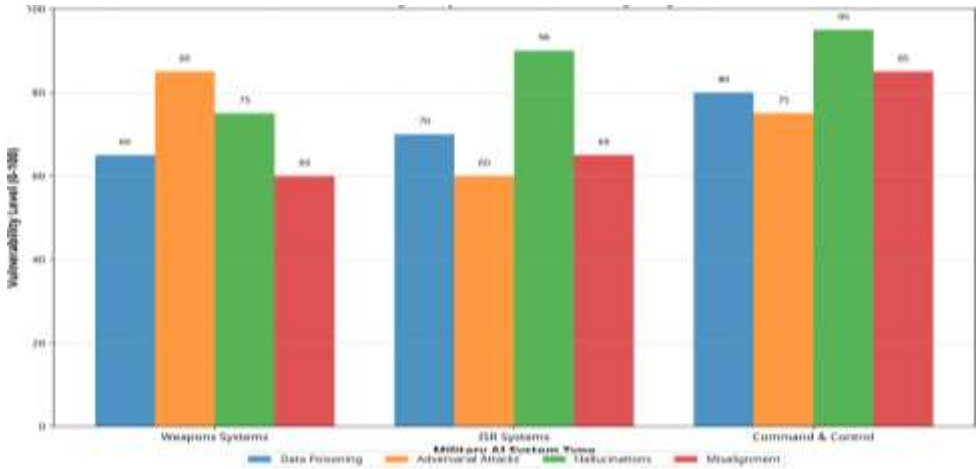

Figure 3: Vulnerability Comparison Across Military AI Systems

Military command and control systems have evolved dramatically in recent decades, from analog communications and

paper maps to sophisticated digital networks that integrate information from multiple domains. The latest evolution involves the incorporation of AI to process vast amounts of data, generate recommendations, and in some cases, automate aspects of the decision cycle. Key AI-enhanced C2 initiatives include:

1. *Joint All-Domain Command and Control (JADC2):* The U.S. military's initiative to connect sensors from all military services into a single network, using AI to process and disseminate data to facilitate faster decision-making.

2. *Advanced Battle Management System (ABMS):* The U.S. Air Force's contribution to JADC2, which uses AI to integrate data from multiple platforms and provide commanders with enhanced situational awareness.

3. *Nuclear Command, Control, and Communications (NC3):* Systems that ensure the command and control of nuclear forces, where AI is being considered for enhancing early warning, threat assessment, and decision support functions.

4. *Automated Decision Support Systems:* AI tools that analyze courses of action and provide recommendations to commanders based on operational data, doctrine, and historical precedents.

These systems aim to accelerate the OODA loop (Observe, Orient, Decide, Act), providing decision advantage in increasingly complex and fast-paced operational environments.

However, the integration of AI into these critical functions also introduces new vulnerabilities, particularly through the potential for hallucinations.

5.2 Critical Vulnerabilities in Decision Support Systems

AI hallucinations in command and control systems can manifest in several ways, each with distinct implications for military operations:

1. *False information generation:* AI systems may confidently present fabricated information as fact, potentially misleading commanders about the operational situation. For example, an AI might hallucinate details about enemy positions, capabilities, or intentions that have no basis in collected intelligence.

2. *Flawed recommendations:* When generating courses of action or recommendations, AI systems may hallucinate potential outcomes, constraints, or opportunities, leading to tactically or strategically unsound suggestions.

3. *Black box problem:* The opacity of many AI systems makes it difficult to understand how they reach their conclusions, potentially obscuring hallucinations behind seemingly authoritative outputs. This

lack of transparency undermines trust and accountability in the command process.

4. *Cyber vulnerability:* AI-enhanced C2 systems may be susceptible to adversarial attacks specifically designed to induce hallucinations, creating a new attack vector for cyber operations.

5. *Misalignment with commander's intent:* AI systems may hallucinate objectives or constraints that don't align with the commander's actual intent, potentially leading to actions that undermine rather than support the mission.

These vulnerabilities are particularly concerning in command and control applications because of the centralized nature of these systems. A hallucination in a C2 system can propagate throughout the force, potentially affecting multiple units and operations simultaneously.

5.3 Special Considerations for Nuclear Command Systems

The potential for AI hallucinations raises particularly grave concerns in the context of nuclear command, control, and communications systems. The stakes in this domain are uniquely high, with the potential for catastrophic consequences from errors or misperceptions. Specific concerns include[5]:

1. *False alarms:* AI systems involved in early warning or threat assessment might hallucinate incoming attacks where none exist, potentially triggering unnecessary escalation or even retaliation.

2. *Compressed decision timelines:* The integration of AI into NC3 systems may accelerate processing speeds beyond human capacity for oversight, creating risks that hallucinations could influence critical decisions before being identified and corrected.

3. *Strategic stability implications:* If multiple nuclear powers integrate AI into their command systems, the potential for hallucination-induced misperceptions could undermine strategic stability and increase the risk of unintended escalation.

4. *Verification challenges:* Unlike conventional military capabilities, it is difficult to verify commitments regarding the role of AI in nuclear command systems, creating potential for misunderstanding and mistrust between nuclear powers.

Given these concerns, many experts and some nuclear-armed states have emphasized the importance of maintaining human control over nuclear weapons decisions. As the U.S. and Chinese leaders

[5] Beyond Human-in-the-Loop: Managing AI Risks in Nuclear Command-and-Control, available at https://warontherocks.com/2024/12/beyond-human-in-the-loop-managing-ai-risks-in-nuclear-command-and-control/, accessed on 05.06.2025.

jointly affirmed in November 2024, there is "the need to maintain human control over the decision to use nuclear weapons." However, even with humans remaining "in the loop," the potential for AI hallucinations to influence human decision-making through false information or misleading recommendations remains a significant concern.

5.4 Human-Machine Interaction Challenges

The effectiveness of AI-enhanced command and control systems depends not only on the technical performance of the AI but also on how humans interact with these systems. AI hallucinations create several challenges in this human-machine relationship:

1. *Automation bias:* Humans tend to trust automated systems, sometimes excessively. This bias may lead commanders to accept AI-generated information or recommendations without sufficient scrutiny, even when they contain hallucinations.
2. *Trust calibration:* Conversely, awareness of the potential for hallucinations may lead to inappropriate distrust of AI systems, potentially negating their benefits even when they are functioning correctly.
3. *Cognitive overload:* The volume and complexity of information in modern military operations already challenges human cognitive capacity. Adding the need to verify AI outputs for potential hallucinations further increases this cognitive burden.
4. *Responsibility and accountability:* When AI systems contribute to the decision- making process, questions arise about responsibility for outcomes, particularly if hallucinations influenced those decisions.
5. *Training and expertise gaps:* Military personnel may lack sufficient understanding of AI limitations, including the potential for hallucinations, hampering their ability to effectively oversee and interpret AI outputs.

Addressing these human-machine interaction challenges requires not only technical solutions but also organizational, doctrinal, and training adaptations to ensure that human operators can effectively leverage AI capabilities while maintaining appropriate oversight.

5.5 Governance Approaches for High-Stakes Decision Systems

Given the significant risks posed by AI hallucinations in command and control systems, robust governance approaches are essential. Drawing lessons from civil nuclear safety regulation, several principles can guide the development of governance frameworks:

1. *Risk-informed governance:* Rather than focusing solely on prescriptive requirements (such as maintaining humans in the loop), governance should quantitatively assess the risks of different system configurations, including the likelihood and consequences of hallucination-induced errors.

2. *Performance-based standards:* Standards should focus on the overall safety and reliability performance of AI-enhanced C2 systems rather than mandating specific technical approaches.

3. *Technology-neutral requirements:* Governance frameworks should establish requirements that can apply across different AI approaches and architectures, allowing for technological evolution while maintaining safety.

4. *Layered defense:* Command and control systems should incorporate multiple, independent layers of protection against hallucination-induced errors, similar to the defense-in-depth approach used in nuclear safety.

5. *Regular assessment and adaptation:* Governance frameworks should include mechanisms for regular assessment of AI performance and adaptation of requirements as technology evolves and new vulnerabilities emerge.

Implementing these governance approaches will require collaboration between military organizations, technical experts, and policy makers to develop standards and protocols that balance the operational benefits of AI integration with the need to mitigate the risks of hallucinations.

As military organizations continue to integrate AI into their command and control systems, addressing the risk of hallucinations must be a central consideration in system design, training, and operational protocols. The potential consequences of command decisions influenced by hallucinated information are simply too severe to ignore, particularly as these systems become more central to military operations across all domains.

5.6 Policy Recommendations

The integration of artificial intelligence into military systems offers significant operational advantages but also introduces new vulnerabilities through the potential for AI hallucinations. Addressing these vulnerabilities requires a comprehensive policy approach that balances innovation with safety, operational effectiveness with reliability, and technological advancement with ethical considerations. The following policy recommendations provide a framework for

military organizations and policymakers to mitigate the risks of AI hallucinations while preserving the benefits of AI integration.

5.7 Ethical Principles for Military AI Development

Even the subject of ethics is quite an old one, now it is more that imperative to deal with and to impose relevant rules and principles. This can be done, for example, by:

1. *Formalizing ethical guidelines:* Military organizations should develop and formalize ethical guidelines specifically addressing AI hallucinations and their potential consequences. These guidelines should emphasize the importance of truthfulness, reliability, and transparency in AI systems.
2. *Prioritizing human well-being:* Ethical frameworks should explicitly prioritize the protection of human life and dignity, recognizing that hallucination-induced errors can have severe humanitarian consequences.
3. *Establishing clear responsibility chains:* Policies should clearly delineate responsibility and accountability for decisions influenced by AI systems, ensuring that appropriate human oversight exists even when hallucinations occur.
4. *Promoting international dialogue:* Military organizations should engage in international dialogue on ethical AI use, working toward shared norms and standards that address the risks of hallucinations in military applications.
5. *Incorporating ethics into acquisition:* Ethical considerations, including the potential for and consequences of hallucinations, should be explicitly incorporated into military AI acquisition processes and requirements.

5.8 Regulatory Frameworks and Compliance Programs

Regulatory frameworks and compliance assurance in this field are mandatory, like:

1. *Establishing AI Security Compliance programs:* Following the model of existing compliance frameworks like PCI for payment security, military organizations should establish AI Security Compliance programs that define standards and best practices for securing AI systems against hallucinations and other vulnerabilities.
2. *Mandate compliance for high-risk applications:* Compliance with these standards should be mandatory for government use of AI systems and for high-risk private sector applications where hallucinations could have severe consequences.
3. *Creating risk-based regulatory tiers:* Regulatory frameworks should establish tiers based on the potential

consequences of hallucinations, with more stringent requirements for systems where hallucinations could lead to loss of life or strategic instability.

4. *Implementing certification processes:* Develop certification processes for military AI systems that include specific testing for hallucination tendencies and mitigation measures.

5. *Establishing oversight bodies:* Create dedicated oversight bodies with appropriate technical expertise to monitor compliance with AI security standards and investigate incidents involving hallucinations.

5.9 Technical Standards and Testing Protocols

Technical aspects of these issues can be managed thru thorough and new relevant standardization processes and testing protocols, possible solutions could resign in:

1. *Developing standardized testing methodologies:* Military organizations should develop standardized methodologies for testing AI systems' susceptibility to hallucinations across different operational conditions and scenarios.

2. *Establishing performance benchmarks:* Define clear, quantitative benchmarks for acceptable hallucination rates in different types of military AI applications, recognizing that zero hallucinations may not be achievable with current technology.

3. *Requiring adversarial testing:* Mandate rigorous adversarial testing of military AI systems to identify vulnerabilities to hallucinations before deployment.

4. *Implementing continuous monitoring:* Develop technical standards for continuous monitoring of deployed AI systems to detect and address hallucinations in operational environments.

5. *Creating shared testing resources:* Establish shared testing environments and datasets that can be used across military organizations to evaluate AI systems' resistance to hallucinations.

5.10 International Cooperation and Treaty Considerations

Some consideration on international cooperation and treaties generation on this subject could follow these aspects:

1. *Pursue confidence-building measures:* Develop international confidence- building measures related to military AI, including information sharing about hallucination mitigation approaches and joint exercises to test interoperability.

2. *Consider arms control frameworks:* Explore the potential for arms control frameworks that address the risks of AI hallucinations in weapons systems, particularly those with

autonomous targeting capabilities.

3. *Establish crisis communication channels:* Create dedicated communication channels between military powers to address incidents or misunderstandings that may arise from AI hallucinations.

4. *Promote transparency in AI capabilities:* Encourage appropriate transparency about military AI capabilities and limitations, including approaches to mitigating hallucinations, to reduce the risk of misperception.

5. *Support international research collaboration*: Foster international collaboration on research to address the technical challenges of AI hallucinations, recognizing this as a shared problem across military organizations.

5.11 Balancing Innovation and Security

A balanced approach between innovation and security is required, among these having the following:

1. *Adopt graduated deployment approaches*: Implement graduated approaches to deploying AI in military systems, beginning with low-risk applications and progressively moving to higher-risk domains as reliability improves.

2. *Establish innovation sandboxes:* Create secure environments where novel military AI applications can be developed and tested for hallucinations without creating operational risks.

3. *Invest in hallucination-resistant AI research:* Allocate significant research funding specifically to developing AI approaches that are more resistant to hallucinations while maintaining performance.

4. *Develop fallback mechanisms:* Require military AI systems to incorporate fallback mechanisms that can detect potential hallucinations and default to safe operational modes.

5. *Balance classification with knowledge sharing:* Find appropriate balances between necessary classification of military AI capabilities and the benefits of knowledge sharing about hallucination risks and mitigation strategies.

5.12 Implementation Strategy

Implementing these policy recommendations will require a coordinated approach across multiple stakeholders:

1. *Military leadership:* Senior military leaders must prioritize addressing AI hallucinations as a critical vulnerability and allocate appropriate resources to mitigation efforts.

2. *Acquisition professionals:* Those responsible for military procurement must incorporate hallucination resistance into

requirements and evaluation criteria for AI systems.

3. *Technical experts:* AI researchers and engineers must work closely with military personnel to develop and implement technical solutions to the hallucination problem.

4. *Training organizations:* Military training programs must be updated to ensure that personnel understand the potential for AI hallucinations and how to identify and respond to them.

5. *International partners:* Collaborative efforts with allies and international organizations will be essential for developing shared approaches to addressing this global challenge.

By implementing these policy recommendations, military organizations can work toward harnessing the benefits of AI integration while mitigating the significant risks posed by hallucinations. This balanced approach recognizes that while hallucinations cannot be completely eliminated with current technology, their risks can be managed through appropriate policies, standards, and practices.

6. CONCLUSION

The integration of artificial intelligence into military systems represents both a transformative opportunity and a significant challenge for armed forces worldwide. As this article has explored, AI hallucinations—instances where AI systems generate outputs that are factually incorrect or entirely fabricated—pose particular risks across weapons systems, intelligence, surveillance, and reconnaissance (ISR) capabilities, and command and control functions. These risks demand serious attention from military leaders, policymakers, and technical experts as AI adoption in military contexts continues to accelerate.

6.1 Summary of Key Findings

Our analysis has revealed several critical insights about the effects of AI hallucinations on military systems:

1. *Inherent vulnerability:* AI hallucinations are not simply bugs or errors that can be eliminated through better programming. They represent inherent limitations in current AI approaches, particularly in machine learning systems that form the backbone of military AI applications. While these limitations may be mitigated, they cannot be completely eliminated with current technology.

2. *Domain-specific manifestations:* Hallucinations manifest differently across military domains, from target misidentification in weapons systems to false pattern recognition in intelligence analysis to flawed recommendations in command and control. Each manifestation carries unique risks that require tailored mitigation

strategies.

3. *Cascading consequences:* The effects of hallucinations can cascade through military systems and processes, potentially leading to serious operational failures, strategic miscalculations, or unintended escalation. This is particularly concerning in high-stakes contexts like nuclear command and control.

4. *Human-machine interaction challenges*: The effectiveness of AI-enhanced military systems depends not only on the technical performance of the AI but also on how humans interact with these systems. Addressing hallucinations requires attention to both technical solutions and human factors, including training, trust calibration, and appropriate oversight.

5. *Governance gaps*: Current regulatory frameworks and military doctrines are not fully equipped to address the unique challenges posed by AI hallucinations. New approaches to governance, testing, and certification are needed to ensure the safe and effective integration of AI into military systems.

6.2 Future Outlook

Looking ahead, several trends will shape the future landscape of AI hallucinations in military systems:

1. *Technical advancements:* Ongoing research in AI safety, explainability, and robustness may yield new approaches that reduce the frequency and severity of hallucinations. However, these advancements are likely to be incremental rather than transformative in the near term.

2. *Increasing autonomy:* The push toward greater autonomy in military systems will continue, driven by operational demands and technological capabilities. This trend will amplify the potential consequences of hallucinations, particularly in systems with lethal capabilities.

3. *Adversarial competition:* As military AI becomes more widespread, adversaries will increasingly develop techniques to deliberately induce hallucinations in opposing systems. This adversarial dimension will add complexity to the hallucination challenge.

4. *Regulatory evolution*: National and international regulatory frameworks for military AI will continue to evolve, with increasing attention to the risks of hallucinations and other AI vulnerabilities. These frameworks will shape how military organizations approach AI integration.

5. *Ethical considerations:* Ethical debates about military AI will increasingly incorporate concerns about hallucinations,

particularly regarding questions of responsibility, accountability, and the potential for unintended harm.

6.3 Final Thoughts on Responsible AI Integration

The challenge of AI hallucinations does not suggest that militaries should abandon AI adoption. Rather, it underscores the need for a thoughtful, measured approach to integration that acknowledges both the potential benefits and risks of these technologies. Responsible AI integration in military contexts requires:

1. *Realistic assessment:* Military planners must realistically assess the capabilities and limitations of AI systems, including their susceptibility to hallucinations, rather than succumbing to hype or fear.

2. *Appropriate human oversight:* While the degree of human involvement may vary across applications, maintaining appropriate human oversight remains essential, particularly for systems where hallucinations could have severe consequences.

3. *Continuous learning:* Military organizations must establish mechanisms for continuous learning about AI performance in operational environments, including systematic tracking and analysis of hallucination incidents.

4. *Ethical framework:* Integration efforts should be guided by robust ethical frameworks that prioritize human well-being, responsibility, and the laws of armed conflict.

5. *International dialogue:* Given the global nature of both AI development and security challenges, international dialogue and cooperation on addressing hallucinations in military AI will be essential.

The effects of AI hallucinations on military systems represent a significant challenge that will require sustained attention from military organizations, policymakers, and technical experts in the years ahead. By acknowledging this challenge and taking proactive steps to address it, we can work toward harnessing the benefits of AI for military applications while mitigating the risks posed by hallucinations. The future of warfare will undoubtedly be shaped by artificial intelligence; ensuring that this future is characterized by reliable, trustworthy, and responsible AI systems must be a priority for all stakeholders in military technology development and deployment.

BIBLIOGRAPHY

COMITER M. (2019). Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It. Belfer Center for Science and International Affairs, Harvard Kennedy School, available at https://www.belfercenter.org/publication/-AttackingAI;

COOK B. (2021). The Future of Artificial Intelligence in ISR Operations. Air University Special Edition Summer 2021, available at https://www.airuniversity.af.edu/Portals/ 10/ASPJ/-journals/Volume-35_Special_Issue/F-Cook.pdf;

SALTINI, A., & Pan, Y. (2024, December 6). Beyond Human-in-the-Loop: Managing AI Risks in Nuclear Command-and-Control. War on the Rocks, available at https:// warontherocks.com/-2024/12/beyond-human-in-the-loop-managing-ai-risks-in-nuclear-command-and-control/;

POMERLEAU M. (2024, October 17). Army's ISR Task Force looking to apply AI to intel data sets. DefenseScoop. available at https://defensescoop.com/2024/10/17/army-isr-task-force-apply-ai-intel-data-sets/;

TOFFOLI J. (2022, June 29). What is Intelligence, Surveillance, and Reconnaissance (ISR) and Why is AI Critical to Military Advantage? Clarifai, available at https:// www.clarifai.com/-blog/what-is-intelligence-surveillance-and-reconnaissance-isr-and- why-is-ai-critical-to-military-advantage;

International Committee of the Red Cross. (2024, September 4). The risks and inefficacies of AI systems in military targeting support. ICRC Blogs, available at https://blogs.icrc.org/law-and-policy/2024/09/04/the-risks-and-inefficacies-of-ai-systems-in-military-targeting-support/;

Foreign Policy in Focus. (2023, July 14). The Military Dangers of AI Are Not Hallucinations. FPIF, available at https://fpif.org/the-military-dangers-of-ai-are-not-hallucinations/;

Google Cloud. (2023). What are AI hallucinations? Google Cloud, available at https://cloud.google.com/discover/what-are-ai-hallucinations;

IBM. (2023). AI hallucinations. IBM Think, available at https://www.ibm.com/think/topics/ai-hallucinations