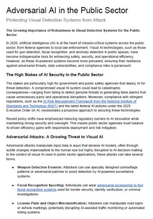
Adversarial AI in the Public Sector

Blog



Adversarial AI in the Public Sector

Protecting Visual Detection Systems from Attack

The Growing Importance of Robustness in Visual Detection Systems for the Public Sector

In 2025, artificial intelligence (AI) is at the heart of mission-critical systems across the public sector, from federal agencies to local law enforcement. Visual AI technologies, such as those used for gun detection, facial recognition, and anomaly detection in public spaces, have become indispensable tools for enhancing safety, security, and operational efficiency. However, as these AI-powered systems become more prevalent, ensuring their resilience against adversarial threats, data vulnerabilities, and compliance risks is paramount.

The High Stakes of AI Security in the Public Sector

The stakes are particularly high for government and public safety agencies that deploy AI for threat detection. A compromised visual AI system could lead to catastrophic consequences—ranging from failing to detect genuine threats to generating false alarms that cause unnecessary panic and operational disruptions. Moreover, compliance with stringent regulations, such as the AI Risk Management Framework from the National Institute of Standards and Technology (NIST) and the latest federal AI policies under the 2025 Executive Order on AI, necessitates a proactive approach to securing these technologies.

Recent policy shifts have emphasized reducing regulatory barriers to AI innovation while maintaining strong security and oversight. This means public sector agencies must balance AI-driven efficiency gains with responsible deployment and risk mitigation.

Adversarial Attacks: A Growing Threat to Visual AI

Adversarial attacks manipulate input data in ways that deceive AI models, often through subtle changes imperceptible to the human eye but highly disruptive to AI decision-making. In the context of visual AI used in public sector applications, these attacks can take several forms:

- **Weapon Detection Evasion:** Attackers can use specially designed camouflage patterns or adversarial patches to avoid detection by AI-powered surveillance systems.
- **Facial Recognition Spoofing:** Individuals can wear adversarial accessories to fool facial recognition systems used for border security, identity verification, or criminal investigations.
- **License Plate and Object Misclassification:** Attackers can manipulate road signs or vehicle markings, potentially disrupting AI-assisted traffic monitoring or automated tolling systems.

Thank you for downloading this AWS and Styrk.AI Blog! Carahsoft is the distributor for AWS public sector solutions available via GSA, NASPO, The Quilt and other contract vehicles.

To learn how to take the next step toward acquiring AWS's solutions, please check out the following resources and information:

For additional resources:
carah.io/AWS-Resources

For upcoming events:
carah.io/AWS-Events

For additional AWS solutions:
carah.io/AWS-Solutions

For additional Marketplace solutions:
carah.io/AWS-Marketplace

To set up a meeting:
AWS@carahsoft.com
888-662-2724

To purchase, check out the contract vehicles available for procurement:
carah.io/AWS-Contracts

For more information, contact Carahsoft or our reseller partners:
AWS@carahsoft.com | 888-662-2724

Adversarial AI in the Public Sector

Protecting Visual Detection Systems from Attack

The Growing Importance of Robustness in Visual Detection Systems for the Public Sector

In 2025, artificial intelligence (AI) is at the heart of mission-critical systems across the public sector, from federal agencies to local law enforcement. Visual AI technologies, such as those used for gun detection, facial recognition, and anomaly detection in public spaces, have become indispensable tools for enhancing safety, security, and operational efficiency. However, as these AI-powered systems become more prevalent, ensuring their resilience against adversarial threats, data vulnerabilities, and compliance risks is paramount.

The High Stakes of AI Security in the Public Sector

The stakes are particularly high for government and public safety agencies that deploy AI for threat detection. A compromised visual AI system could lead to catastrophic consequences—ranging from failing to detect genuine threats to generating false alarms that cause unnecessary panic and operational disruptions. Moreover, compliance with stringent regulations, such as the AI Risk Management Framework from the National Institute of Standards and Technology (NIST) and the latest federal AI policies under the 2025 Executive Order on AI, necessitates a proactive approach to securing these technologies.

Recent policy shifts have emphasized reducing regulatory barriers to AI innovation while maintaining strong security and oversight. This means public sector agencies must balance AI-driven efficiency gains with responsible deployment and risk mitigation.

Adversarial Attacks: A Growing Threat to Visual AI

Adversarial attacks manipulate input data in ways that deceive AI models, often through subtle changes imperceptible to the human eye but highly disruptive to AI decision-making. In the context of visual AI used in public sector applications, these attacks can take several forms:

- **Weapon Detection Evasion:** Attackers can use specially designed camouflage patterns or adversarial patches to avoid detection by AI-powered surveillance systems.

- **Facial Recognition Spoofing:** Individuals can wear adversarial accessories to fool facial recognition systems used for border security, identity verification, or criminal investigations.

- **License Plate and Object Misclassification:** Attackers can manipulate road signs or vehicle markings, potentially disrupting AI-assisted traffic monitoring or automated tolling systems.

Research has demonstrated that adversarial perturbations can significantly degrade the accuracy of image classification models, with targeted attacks achieving success rates of 70% to 90% or higher, depending on the model and attack type. These attacks involve subtle modifications to inputs that are imperceptible to humans but profoundly impact model decision-making. This level of vulnerability is unacceptable for public safety applications, such as image classification, where high reliability and accuracy are non-negotiable.

Why Public Sector Visual AI Systems Are Particularly Vulnerable

Government agencies face unique challenges when deploying visual AI systems, including:

- **Operational Environment Variability:** Public sector AI must function across diverse real-world conditions, including varying lighting, weather, and urban infrastructure.
- **High-Value Targets for Attackers:** AI-driven security systems in airports, critical infrastructure, and law enforcement settings are prime targets for adversaries looking to exploit system weaknesses.
- **Procurement and Legacy System Integration:** Many public-sector AI systems must integrate with legacy security infrastructure, which may introduce additional vulnerabilities.

The Consequences of Adversarial Attacks on Public Sector AI

For mission-critical applications such as gun detection in schools or public transit security monitoring, the consequences of adversarial attacks include:

- **False Positives:** Misclassifications could lead to unnecessary lockdowns, wasted law enforcement resources, and public distrust in AI security measures.
    - *Example: In September 2024, New York City piloted an AI-enabled weapons scanner in its subway system, which produced 118 false positives, leading to unnecessary searches.* (AI Incident Database).
- **False Negatives:** Failure to detect a genuine threat due to adversarial manipulation could have severe consequences, including loss of life.
    - *Example:* Israel's military has integrated AI technology from U.S. companies to enhance target identification efficiency. However, failures to accurately identify legitimate threats have led to civilian casualties. (AP News).
- **Bias:** Adversarial attacks could exacerbate existing biases within AI systems, leading to disproportionate targeting or misclassification of certain groups.
    - *Example:* Facial recognition systems used by law enforcement have shown higher error rates for people with darker skin tones and females, potentially

leading to wrongful investigations and arrests. (PubMed).

How Public Sector Agencies Can Protect Visual AI Systems

To safeguard mission-critical visual AI systems, government agencies must adopt a proactive, security-first approach. Best practices include:

1. Implementing AISecOps for Government AI Deployments

AISecOps (AI Security Operations) integrates security into every stage of the AI lifecycle, ensuring that vulnerabilities are identified and mitigated before deployment. Agencies should:

- Conduct red teaming exercises to test AI model robustness.
- Develop adversarially trained models to enhance resilience.
- Ensure ongoing monitoring for adversarial activity and model drift.

2. Strengthening Regulatory Compliance and AI Governance

With the 2025 AI Executive Order emphasizing both innovation and security, agencies should align with updated federal guidelines and frameworks such as NIST's AI RMF. Key actions include:

- Establishing cross-agency collaboration on AI security policies.
- Ensuring transparency and explainability in AI decision-making.
- Regularly auditing AI models for bias and adversarial weaknesses.

3. Enhancing Training Data and Model Hardening Techniques

- Utilize diverse and adversarially augmented datasets to improve model robustness and reduce susceptibility to adversarial attacks.
- Conduct vulnerability testing to identify and address adversarial weaknesses, ensuring models can withstand various attack strategies.
- Employ ensemble methods and hybrid AI architectures to mitigate single-point failures and enhance resilience.
- Implement defense techniques such as adversarial training, gradient masking, and robust optimization to harden models against potential attacks.

- Establish real-time monitoring systems to detect anomalies in AI outputs and trigger immediate responses to mitigate the impact of attacks.

4. Leveraging Public-Private Partnerships for AI Security

Collaboration between federal agencies, AI research institutions, and cybersecurity firms can drive innovation in securing AI systems. For example, the Department of Homeland Security (DHS) has partnered with AI vendors to test and improve the resilience of AI-powered border security solutions.

Building Trust in Public Sector AI Systems

Ensuring the security and reliability of AI in government applications is not just a technical challenge—it's a public trust issue. Public sector agencies must be proactive in addressing adversarial threats and compliance challenges to maintain confidence in AI-powered security solutions.

At Styrk AI, we specialize in helping government agencies and organizations secure their AI/ML models against adversarial threats and regulatory risks. Our solutions include:

- **AI Risk Assessment & Management:** Identify and mitigate vulnerabilities in public sector AI applications.

- **Model Hardening for Government Use Cases:** Strengthen AI models against adversarial attacks and emerging threats.

- **AISecOps Framework Adoption:** Ensure security-by-design practices are embedded in government AI deployments.

As AI continues to shape the future of public safety, securing visual detection systems must remain a top priority. Agencies that take a proactive approach to AI security will be better positioned to harness AI's full potential while safeguarding citizens and critical infrastructure.

To learn more about how Styrk AI helps the public sector uncover and defend vulnerabilities in visual AI/ML models, visit our website.

Anhad Singh | Founder & CEO, Styrk AI
Anhad is currently CEO at Styrk AI, and has a PhD in Data Privacy from University of California, Davis. Before starting Styrk AI, Anhad led multiple internal and external privacy initiatives at Google as Sr. Privacy Engineer. Prior to Google, he was at Dataguise (acquired in 2020), leading multiple responsibilities across multiple roles, including designing and building software, Product Management and the Head of Technology and Partnerships. Anhad loves to work at the intersection of technology, business and legal, which stems from his passion to create a world of secure and trustworthy systems.