Realizable Attacks: When Adversarial AI Meets the Real World

March 14, 2025

1 Introduction

In the evolving landscape of artificial intelligence and cybersecurity, a particularly concerning threat has emerged at the intersection of theoretical research and practical application: realizable attacks. Unlike purely theoretical adversarial attacks that may work only in controlled laboratory conditions, realizable attacks—also known as problem-space attacks—successfully bridge the gap between digital vulnerabilities and physical reality, creating tangible threats to AI systems deployed in the real world.

Realizable attacks represent a sophisticated evolution in the adversarial machine learning field. They go beyond simply fooling a model within a digital environment to creating adversarial examples that remain effective when implemented in real-world contexts, accounting for physical constraints, environmental variables, and system preprocessing. As Pierazzi et al. (2020) articulate in their seminal work, "Realizable attacks require the adversary to reason about the inverse feature-mapping problem: how to modify objects in the problem space so that their feature space representation changes in a way that induces misclassification."

The significance of these attacks extends far beyond academic interest. As AI systems increasingly control critical infrastructure and decision-making processes, the ability to manipulate these systems in real-world settings poses serious risks to public safety, privacy, and security. Self-driving vehicles could misinterpret traffic signs with subtle modifications, biometric authentication systems might fail to distinguish between legitimate users and attackers using specially crafted inputs, and fraud detection systems could be rendered ineffective against transactions specifically designed to appear legitimate while concealing fraudulent activity.

The danger is particularly acute because many organizations deploy AI models with strong performance in controlled settings but insufficient testing against adversarial inputs that account for real-world constraints and implementation factors. This gap between theoretical and practical security creates a vulnerability landscape that sophisticated attackers can exploit with potentially severe consequences.

2 Understanding Realizable Attacks

Realizable attacks distinguish themselves from theoretical adversarial attacks by accounting for the practical constraints of the real world. While a theoretical attack might suggest pixel-level changes to an image that would fool an AI classifier, a realizable attack must consider whether those changes can be implemented in a physical object or environment and still retain their effectiveness when captured by sensors, processed by various algorithms, and finally analyzed by the target model.

As Cavallaro and De Cristofaro (2023) explain in the Security and Privacy of AI Knowledge Guide, "Realizable attacks are a category of adversarial attacks that focus on creating real adversarial objects." They emphasize that "for instance, in malware classification tasks, realizable adversarial attacks are not

just concerned at creating digital adversarial examples...but also focused at generating an adversarial object that exists in the physical world."

The key difference lies in what security researchers call the "problem space" versus the "feature space." The problem space encompasses the actual physical or digital objects (such as images, malware code, or audio signals) as they exist in reality. The feature space, by contrast, represents the mathematical abstraction of these objects that machine learning models operate on. Realizable attacks must work within the constraints of the problem space while achieving the desired effect in the feature space.

Consider a facial recognition system used for secure access. A theoretical attack might suggest specific pixel modifications to a photograph to fool the system. However, a realizable attack needs to consider:

1. How to translate those pixel modifications into physical changes (makeup, accessories, lighting conditions)
2. Whether those physical changes will be captured consistently by cameras in different lighting and angles
3. If preprocessing steps like normalization or alignment might neutralize the attack
4. Whether the attack remains effective after compression or transmission through the system

Real-world examples highlight these challenges. In autonomous driving contexts, attackers have demonstrated that physical modifications to road signs—such as strategically placed stickers—can cause vision systems to misclassify them entirely. Similarly, researchers have shown how specially designed eyeglass frames can defeat facial recognition systems consistently in the real world, despite variable lighting, angles, and distances (Sharif et al., 2016).

3 How Realizable Attacks Work

The technical mechanisms behind realizable attacks involve sophisticated understanding of both AI systems and the physical world. Several key approaches characterize how these attacks operate:

3.1 Adversarial Perturbations with Real-World Constraints

Realizable attacks typically begin with traditional adversarial perturbations—carefully calculated modifications to inputs that cause AI systems to make incorrect predictions. However, they add a crucial layer of constraint: the perturbations must be implementable in the physical world.

For example, when creating an adversarial attack against an image recognition system, the attacker must consider:

- **Physical realizability**: Changes must be physically implementable (e.g., using stickers, paint, or lighting)
- **Environmental invariance**: The attack should work across different lighting conditions, angles, and distances
- **Transformation robustness**: The attack should remain effective after the physical object is captured by sensors and processed by the system

As noted by Eykholt et al. (2018) in their work on robust physical-world attacks on deep learning visual classification, "Perturbations are usually measured with Lp norms, but these mathematical distance metrics may not represent actual concern for safety-critical systems." They emphasize that physical realizability requires optimizing for "non-suspiciousness and robustness to environmental conditions."

3.2 Side-Effect Features and Semantic Preservation

A critical aspect of realizable attacks is managing what researchers call "side-effect features." When making changes to objects in the real world, those modifications often have unintended consequences or side effects on other features that the model might analyze.

For instance, when modifying malware to evade detection, simply changing certain byte sequences might render the malware non-functional. A realizable attack must therefore maintain the semantic functionality of the malware while altering its feature representation enough to evade detection. Similarly, physical modifications to objects must preserve their original purpose and appearance to humans while fooling machines.

Pierazzi et al. (2020) describe this challenge: "Projecting adversarial points from the feature space back to the problem space introduces side-effect features as a byproduct of satisfying problem space constraints." They note that these side-effect features "exist to make the attack realistic, as they facilitate adherence to the inherent constraints of the problem space."

3.3 Transformation-Aware Optimization

Realizable attacks must account for the various transformations that occur when physical objects are processed by computational systems. This includes:

1. **Sensor variability**: Different cameras, microphones, or sensors might capture the same physical object differently
2. **Preprocessing algorithms**: Systems often normalize, crop, or filter inputs before analysis
3. **Feature extraction**: The conversion from raw sensory data to feature vectors can affect attack success

Sophisticated realizable attacks therefore use transformation-aware optimization techniques, often involving end-to-end simulations of the entire pipeline from physical object to final classification. The attacker optimizes the perturbation not just for a single digital input but for robustness across the entire process chain.

For example, when attacking facial recognition, researchers might simulate different camera angles, lighting conditions, and preprocessing steps to ensure their physical adversarial examples (like specially designed glasses) work consistently in varied real-world conditions.

4 Categories of Realizable Attacks

Realizable attacks span multiple domains and techniques, each tailored to specific AI applications and their real-world deployment contexts.

4.1 Physical Adversarial Examples

Physical adversarial examples represent perhaps the most intuitive category of realizable attacks. These involve creating physical objects specifically designed to fool AI classifiers. Notable examples include:

- **Traffic sign attacks**: Researchers have demonstrated how carefully placed stickers on stop signs can cause them to be misclassified as speed limit signs by autonomous vehicle vision systems (Eykholt et al., 2018).

- **3D printed objects**: Specially designed 3D objects can be consistently misclassified regardless of viewing angle or lighting conditions.
- **Adversarial patches**: Wearable patches or stickers that can cause person detection systems to fail to recognize humans in the scene.

The key challenge in this category is creating modifications that remain effective across different distances, angles, and lighting conditions—all while appearing inconspicuous to humans.

4.2 Transformation-Robust Digital Attacks

These attacks target digital systems but account for real-world preprocessing and transformation steps. Examples include:

- **Robust audio adversarial examples**: Audio commands designed to control voice assistants even after being played through speakers and captured by microphones, potentially with background noise.
- **Camera-robust adversarial images**: Digital images crafted to remain adversarial even after being displayed on screens, captured by cameras, and processed by computer vision systems.

These attacks must account for quality loss, format conversion, and various transformations that occur in real-world digital pipelines.

4.3 Sensor-Based Attacks

Sensor-based attacks specifically target the sensing mechanisms that AI systems use to perceive the world. Examples include:

- **LiDAR spoofing**: Creating false returns in autonomous vehicle LiDAR sensors to generate phantom obstacles.
- **Microphone jamming**: Using ultrasonic signals to inject inaudible commands into voice recognition systems.
- **Camera blinding**: Using precisely timed light pulses to temporarily blind camera-based systems at critical moments.

These attacks exploit physical properties of sensors rather than just the machine learning models that process their data.

4.4 Semantically Constrained Attacks

These attacks operate in domains where the adversarial inputs must maintain specific semantic properties to remain functional. Examples include:

- **Functional malware**: Malware that evades detection while preserving its malicious functionality.
- **Adversarial text**: Text that maintains grammatical correctness and semantic meaning to humans but causes NLP systems to make incorrect classifications.
- **Executable adversarial examples**: Code modifications that preserve program behavior while evading analysis tools.

These attacks are particularly challenging as they must balance the competing objectives of evading detection and maintaining functionality.

5 Where & When Realizable Attacks Are Used

Realizable attacks find application across numerous domains where AI systems interface with the physical world or must process inputs from potentially adversarial sources.

5.1 Autonomous Vehicle Security

Self-driving vehicles represent a prime target for realizable attacks due to their reliance on various sensors and AI vision systems for critical safety decisions. Attackers could potentially:

- Modify road signs with carefully placed stickers to cause misclassification
- Create adversarial patterns on vehicles or buildings that make them invisible to object detection systems
- Deploy physical objects designed to be misclassified as different objects (e.g., making a stop sign appear as a yield sign)

As noted by Moreno-Torres et al. (2012), "These attacks are particularly dangerous because they can cause systems to make incorrect decisions in safety-critical scenarios." The consequences could range from unauthorized access to restricted areas to potentially catastrophic accidents.

5.2 Biometric Security Circumvention

Biometric authentication systems increasingly rely on AI for facial recognition, fingerprint matching, and voice identification. Realizable attacks in this domain include:

- Creating physical masks or accessories that fool facial recognition into identifying the attacker as an authorized user
- Developing artificial fingerprints that match specific individuals
- Synthesizing voice samples that bypass voice recognition authentication

These attacks are particularly concerning because biometric systems are increasingly used for high-security applications like banking, border control, and facility access.

5.3 Evading AI-Powered Threat Detection

In cybersecurity, numerous tools now employ AI to detect malware, phishing attempts, and intrusions. Realizable attacks in this context include:

- Creating malware that maintains its functionality while evading AI-based detection
- Designing phishing websites that appear legitimate to both humans and AI safety tools
- Crafting network traffic patterns that avoid triggering anomaly detection systems

These attacks directly undermine security infrastructure designed to protect organizations and individuals from cyber threats.

5.4 Manipulating Financial and Fraud Detection Systems

Financial institutions heavily rely on AI to detect fraudulent transactions and suspicious activities. Attackers can design realizable attacks that:

- Structure financial transactions to avoid triggering fraud detection algorithms
- Create fake documents that pass automated verification systems
- Generate synthetic identities that appear legitimate to KYC (Know Your Customer) AI systems

The financial incentives for such attacks are substantial, making this an active area for both attackers and defenders.

6 Real-World Case Studies

6.1 Case Study 1: The Turtle Mistaken for a Rifle

In 2019, researchers from MIT demonstrated a striking example of a realizable attack by 3D-printing a turtle with a specific texture pattern. When viewed by advanced object recognition systems, the turtle was consistently misclassified as a rifle, regardless of the viewing angle or lighting conditions (Athalye et al., 2018).

This case highlighted several critical insights:

1. **Physical robustness**: The attack remained effective across various physical transformations, including rotation, lighting changes, and different viewpoints.
2. **Inconspicuousness**: To humans, the object clearly looked like a turtle, with the adversarial pattern appearing merely as an interesting texture.
3. **Transferability**: The attack worked against multiple different vision models, not just the one it was optimized for.

The implications were significant—demonstrating that physical objects could be manufactured specifically to fool AI systems consistently in real-world conditions. This raised concerns about how such techniques might be used to evade security systems or fool autonomous vehicles.

6.2 Case Study 2: Evading Face Recognition with Adversarial Eyeglasses

Sharif et al. (2016) demonstrated a particularly concerning realizable attack against facial recognition systems. They created specially designed eyeglass frames that, when worn, could cause the wearer to be misidentified as someone else or to evade identification altogether.

The attack involved:

1. Printing the eyeglass frames with a precisely calculated pattern
2. Testing across different lighting conditions and camera angles
3. Ensuring the attack remained effective despite various preprocessing steps in facial recognition systems

The researchers achieved up to 100% success rates in targeted impersonation attacks against state-of-the-art facial recognition systems. What made this attack particularly notable was:

- **Practicality**: The eyeglasses could be easily manufactured and worn
- **Inconspicuousness**: They appeared as normal fashion accessories
- **Effectiveness**: They worked consistently in real-world conditions
- **Specificity**: They could target impersonation of specific individuals

This demonstration raised serious concerns about the security of facial recognition systems used for access control and surveillance, showing how physical accessories could fundamentally undermine their reliability.

7 Defensive Strategies Against Realizable Attacks

As the threat of realizable attacks grows, researchers and organizations have developed various defensive strategies to enhance AI system robustness against these sophisticated attacks.

7.1 Robust Data Augmentation and Training

One of the most effective approaches involves training AI systems on data that reflects real-world variations and potential adversarial manipulations:

- **Adversarial training**: Intentionally including adversarial examples in training data
- **Domain randomization**: Training models with highly varied backgrounds, lighting, and perspectives
- **Physical transformation simulation**: Incorporating simulations of real-world physical transformations (rotation, lighting changes, sensor noise) during training

As demonstrated by RealizableAttack (2023) in their implementation: "With data augmentation, the model learns to recognize digits across various transforms, making it more robust against realizable adversarial attacks that use similar transformations."

7.2 Multi-sensor Fusion and Cross-Validation

Systems that rely on multiple independent sensors and cross-validate their inputs are significantly harder to attack:

- **Sensor diversity**: Using different types of sensors (cameras, LiDAR, radar, infrared) to perceive the same environment
- **Cross-validation**: Requiring consistent object recognition across multiple sensing modalities
- **Temporal consistency checking**: Verifying that detections remain consistent over time

When one sensor type might be vulnerable to a particular attack, others with different physical properties may remain unaffected, providing a defense-in-depth approach.

7.3 Anomaly Detection for Adversarial Inputs

Specialized systems can be deployed to detect inputs that have characteristics of adversarial examples:

- **Statistical analysis**: Flagging inputs with unusual statistical properties
- **Perturbation detection**: Identifying patterns characteristic of adversarial perturbations
- **Out-of-distribution detection**: Recognizing inputs that differ significantly from typical examples

As noted by Carlini and Wagner (2017), "Detecting adversarial examples is a promising approach to defend against attacks, but detectors must be robust to adaptive attackers who know about the detection mechanism."

7.4 Security by Design in AI Systems

Fundamentally secure AI deployment requires considering security throughout the design process:

- **Interpretable models**: Using models that provide explanations for their decisions, making it easier to identify suspicious inputs
- **Formal verification**: Mathematically proving properties about model behavior under certain conditions
- **Secure model deployment**: Implementing secure update mechanisms, monitoring, and response protocols

Organizations should incorporate adversarial testing into their development lifecycle, regularly testing systems against state-of-the-art realizable attacks.

8 Conclusion

Realizable attacks represent a crucial evolution in the field of adversarial machine learning—bridging the gap between theoretical vulnerabilities and practical exploits. As AI systems increasingly make critical decisions in our physical world, the ability of attackers to craft inputs that function effectively in real-world conditions poses significant security, safety, and privacy concerns.

The technical sophistication of these attacks continues to grow, with researchers and malicious actors developing increasingly robust methods to fool AI systems while accounting for real-world constraints. From traffic signs modified to mislead autonomous vehicles to accessories designed to defeat facial recognition, these attacks exploit the fundamental gap between the simplified mathematical representations used by AI systems and the complex, noisy reality they attempt to interpret.

Defending against such attacks requires a multi-faceted approach. Organizations must combine robust training techniques, diverse sensing modalities, anomaly detection, and security-by-design principles to create AI systems resistant to realizable attacks. Equally important is acknowledging that perfect security is unattainable—continuous testing, monitoring, and improvement are essential in the face of evolving threats.

As we continue to integrate AI into critical infrastructure and decision-making processes, understanding and addressing realizable attacks becomes not merely an academic exercise but a practical necessity for responsible deployment. The challenge for security professionals and AI developers is clear: design systems that maintain their integrity even when faced with carefully crafted inputs specifically designed to manipulate them in the messy, complex reality of the physical world.

References

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. (2018). Synthesizing robust adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, pages 284–293.

Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57.

Cavallaro, L. and De Cristofaro, E. (2023). Security and privacy of ai: Knowledge guide. *CyBOK*.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Computer Vision and Pattern Recognition (CVPR)*.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530.

Pierazzi, F., Pendlebury, F., Cortellazzi, J., and Cavallaro, L. (2020). Intriguing properties of adversarial ml attacks in the problem space. In *IEEE Symposium on Security and Privacy (SP)*, pages 1308–1325.

RealizableAttack (2023). Realizable (problem-space) attacks on mnist classifier. CyBOK AI Lab.

Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540.