AI Red Teaming:
Facial Recognition Case Study

ADVERSA

Speaker: Alex Polyakov

- 18 years in Cybersecurity, 5 years in AI
- Founder: Adversa.AI
- Member: Forbes Technology Council
- Author: 2 books, first AI Security MOOC
- Speaker: 100+ conferences in 30+ countries
- Hobbies: SynthBio, Neuroscience, PsyTech

About Adversa.AI

Startup with a mission to increase trust in AI systems

by protecting them from cyber threats

AI Risk Advisory (Awareness)

Understanding threats to AI systems

- **Report** on Adversarial ML history
- **Course** for Adversarial ML practice
- **Newsletter** on vulnerabilities and incidents

AI Red Team (Assessment)

Finding vulnerabilities in AI systems

- **Vulnerability** research in AI systems
- **Speaking** at conferences on AI security
- **AI red teaming** and security auditing

AI Security Engineering (Assurance)

Implementing defenses for AI systems

- **Mitigation** research for AI systems
- **Framework** Adversarial Octopus
- **Sample lifecycle** for Secure AI

Agenda

- Secure AI 101
- AI Red Teaming
- Digital attacks
- Physical attacks
- Defenses
- Takeaways

Secure AI 101

Why Securing AI is Essential

Traditional Software	AI Systems
Powered by Fixed program logic	**Powered by** Flexible ML training
Interaction Graphical UI: menus and buttons Workflow: tasks and commands	**Interaction** Cognitive UI: vision, audition, natural language Workflow: learning and decisions
Typical attacks Improper validation Access control issues Security misconfiguration	**Typical attacks** Evasion Poisoning Data exfiltration

*Threat landscape is changing!

Why Securing AI

Confidentiality

- *Personal data is extracted from Netflix statistics shared for ML contest*

Integrity

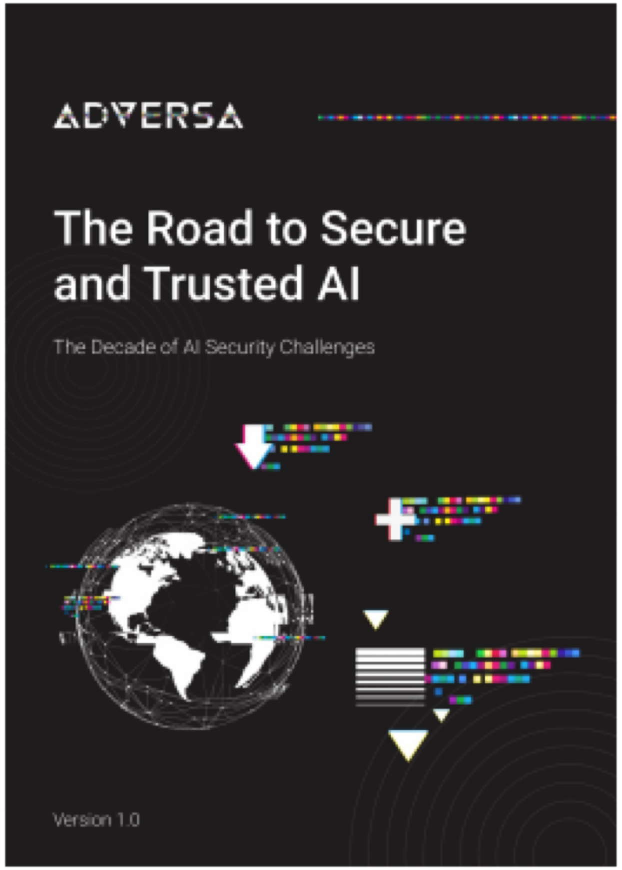
- *Malware bypasses an AI-based threat detection from Blackberry/Cylance*

Availability

- *Self-driving autopilot keeps causing Tesla cars' crashes*

What We are Talking About (Applications)

Attacked AI applications	Share
Image classification	43.4%
Face recognition	6.7%
Data analytics	6.4%
Malware detection	4.3%
Speech recognition	3.0%
Sentiment analysis	2.9%
Object detection	2.7%
Reinforcement learning	2.7%
Semantic segmentation	2.0%
Medical imaging	1.9%
Other 48 applications	*24.1%*



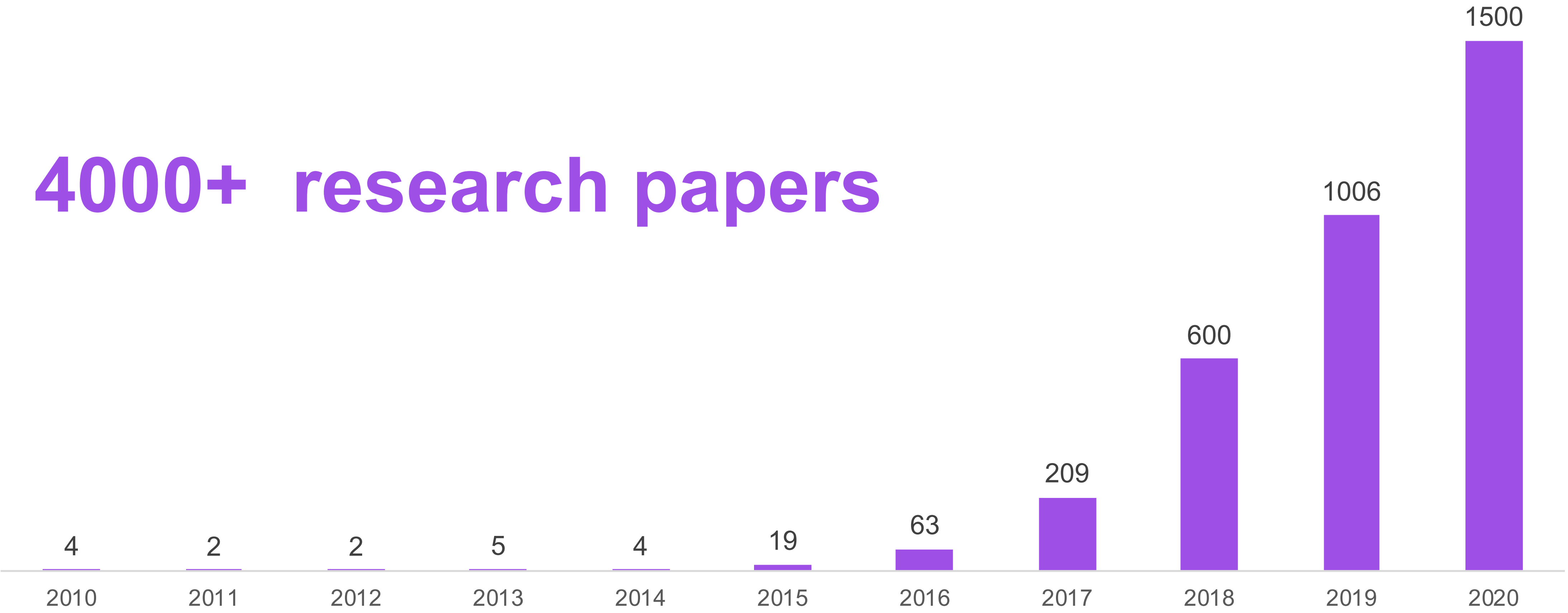
Source: Adversa.AI "The road to secure and Trusted AI" Whitepaper

Who is Affected (AI-Powered Industries)

- **Automotive**
  - Object detection, Semantic segmentation, **Facial recognition**, …
- **Biometrics**
  - Behavior analysis, Speaker verification, **Facial recognition**, …
- **Smart home**
  - Speech recognition, Question answering**, Facial recognition,** …
- **Robotics**
  - Reinforcement learning, Action recognition, **Facial recognition**,
- **Healthcare**
  - Medical imaging, Electrodiagnosis, **Facial recognition**, …
- **Internet**
  - Sentiment analysis, Content moderation, **Facial recognition**, …
- **Retail**
  - Data analytics, Image classification, **Facial recognitio**n, …
- **Finance**
  - Text analytics, Event prediction, **Facial recognition**, …
- **Industry 4.0**
  - Predictive maintenance, System profiling, **Facial Recognition** …
- **Cybersecurity**
  - Threat detection, System profiling, Malware , **Facial Recognition**

When it was started (History)

4000+ research papers



Source: Adversa.AI "The road to secure and Trusted AI" Whitepaper

How to Attack (Top 10 Attacks)

1. Evasion attack
bypasses normal decisions by AI systems in favor of attacker-controlled behavior by crafting malicious data inputs called adversarial examples

2. Poisoning attack
reduces the quality of AI decisions while making AI systems unreliable or unusable by injecting malicious data into a dataset used for AI training

3. Membership inference attack
discloses whether specific data sample was a part of a dataset used for AI training

4. Backdoor attack
invokes hidden behavior of AI systems after poisoning them with secret triggers while keeping AI models work as intended in normal conditions

5. Model extraction attack
exposes algorithm’s internal details by making malicious queries to AI systems

6. Attribute inference attack
reveals secret data details by exploiting public information received from AI system responses

7. Trojan attack enables
attacker-controlled behavior of AI systems after malicious modification or distribution of AI models that work as expected in normal conditions

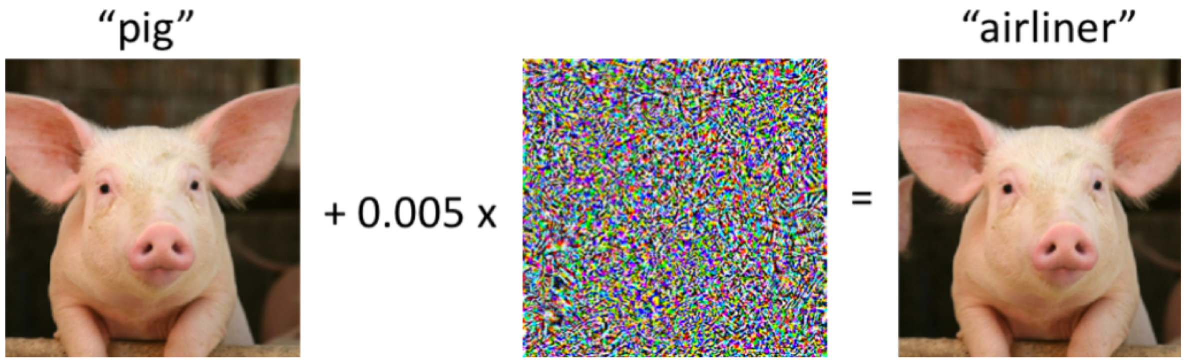
8. Model inversion attack
reveals secret data inputs based on public outputs by maliciously querying AI systems

9. Anti-watermarking attack
bypasses protection controls used by AI systems for copyright or authenticity checks

10. Reprogramming attack
allows attackers to repurpose AI models and make them execute new tasks

Evasion Attacks in Computer Vision

- Discover the most important pixels by interacting with a model
- Craft a malicious image with modified pixels to fool a model
- The model makes a wrong prediction controlled by an attacker
- Changes are imperceptible for system owners



Academic AI Attacks Against Real Software

- Internet companies
- Cybersecurity software
- Facial recognition services
- Identity verification vendors
- Self-driving cars, such as Tesla
- Content moderation on Facebook or Twitter
- Copyright detection, such as YouTube Content ID
- Speech recognition, such as Alexa, Siri, Mozilla, etc.
- Specialized AI platforms, such as Clarifai or BigML etc.
- General cloud platforms, such as Google, Microsoft, Amazon, IBM, etc.

Real AI Attacks by Threat Actors

- Infection of Microsoft’s Tay bot
- Infection of VirusTotal database
- Evasion of spam detection systems
- Evasion of copyright protection systems
- Evasion of behavioral ML-based AV detection
- Evasion of content moderation on Twitter by China
- Manipulation of news feeds and search results (black SEO)
- Manipulation of exam auto-grading NLP systems by students
- Manipulation of HR auto-screening NLP systems by applicants
- Manipulation of Chinese national facial recognition for tax fraud
- Exfiltration of personal data used for AI system training in Korea
- Exfiltration of commercial data and software used by Clearview AI

Growing AI Red Teams

Types of AI red teams

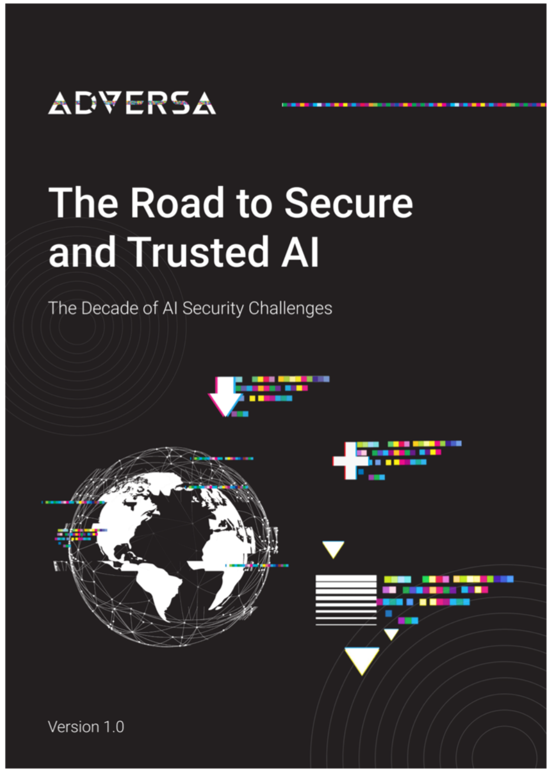
- *Regular security researchers exploring AI security*
- *Specialized AI red teams testing utilized AI systems*
- *Consulting AI red teams offering AI security advisory*

Example AI red teams

- *Adversa.AI, Bosch, DoD, Facebook, Microsoft, MITRE, Nvidia, OpenAI, etc*

Read More on Adversa.AI’s Report

- The demand for secure AI
- The inception of trustworthy AI
- Progress toward secure AI
- Why security of AI matters
- What AI areas are in danger
- Who is in danger exactly
- When AI security started
- Where AI risks are addressed
- How AI attacks and defenses work
- Lifecycle for security of AI



URL: https://adversa.ai/hitb

AI Red Teaming Scenario

How to start attack testing

ADVERSA

Problem

- Over 4000 research papers in adversarial ML
- Countless combos of different conditions
- No clear understanding of real-world risks beyond research
- We proposed our own approach

Choose Attack Goals

- **Confidence reduction** – make results unusable
  - *Decrease a confidence score without changing a prediction*
- **Misclassification** – avoid person detection
  - *Change a prediction from the expected class to random one*
- **Targeted misclassification** – pretend to be another person
  - *Change a prediction to the desired class*

Choose Attack Forms

- Glasses
- Lenses
- Mask
- Hat
- Moustache
- Band-aid
- Etc.

Choose Attack Actor

- Whitebox testing (Attacker have some access to model)
- Greybox testing (Attacker have some access to API)
- Blackbox testing (Attacker have some access to device)

Choose Attack Conditions

- Digital environment (2d)
  - *Adjust to various preprocessing (compression, clipping)*
- Physical environment (3d)
  - *Adjust to printing issues (size, color inconsistency, position)*
- Dynamic physical environment (4d+)
  - *Adjust to various patch positions (sizes, angles, etc.)*

Choose Attack Methods

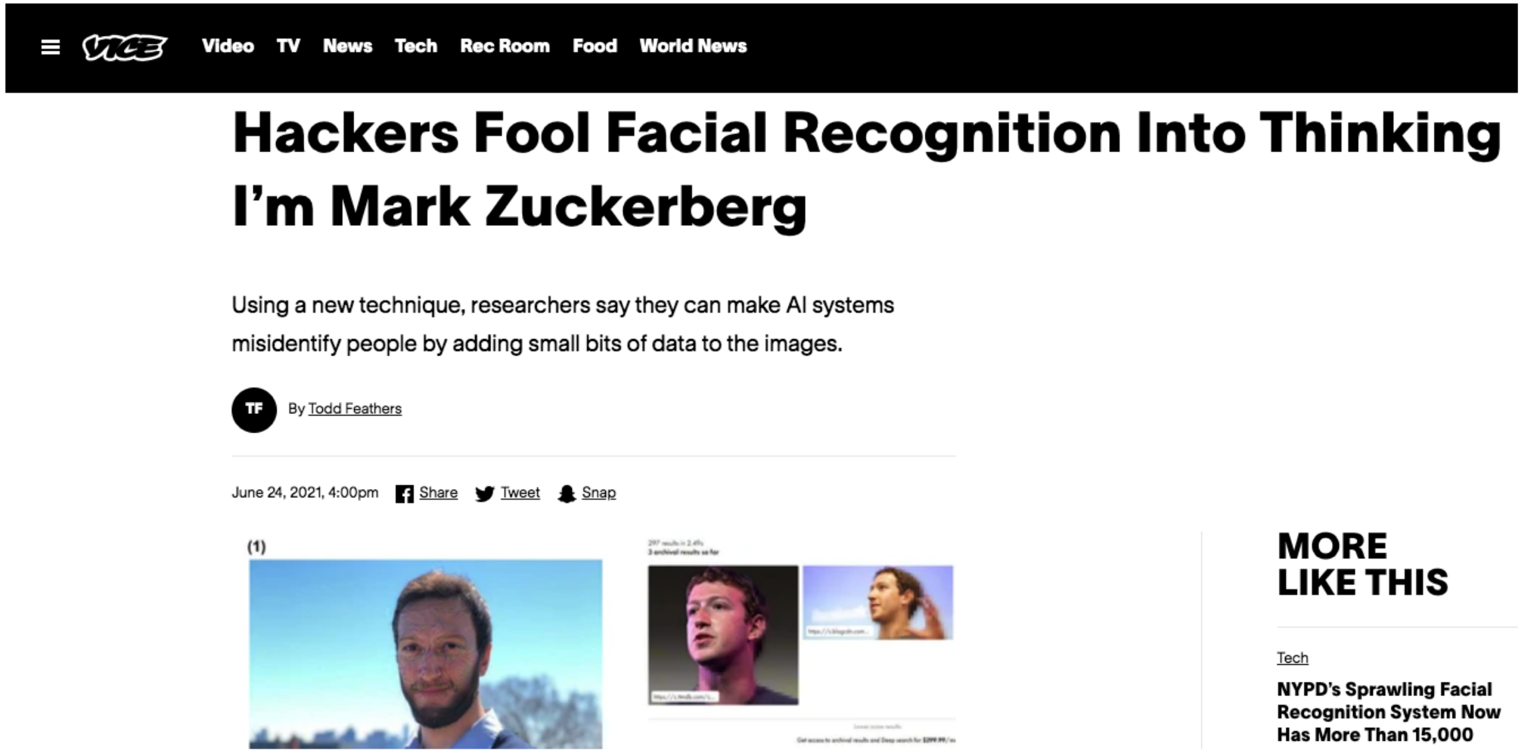
- FGSM
- BIM
- PGD
- EoT
- DeepFool

+Hundreds of various other methods

Attack Success Criteria

- **Misclassification** – attack success rate
- **Imperceptibility** – difficulty to detect a malicious input by humans
- **Transferability** – attack stability against changing environments

Digital Attack Demo (PimEyes.com)



Video TV News Tech Rec Room Food World News

Hackers Fool Facial Recognition Into Thinking I'm Mark Zuckerberg

Using a new technique, researchers say they can make AI systems misidentify people by adding small bits of data to the images.

By Todd Feathers

June 24, 2021, 4:00pm Share Tweet Snap

(1)

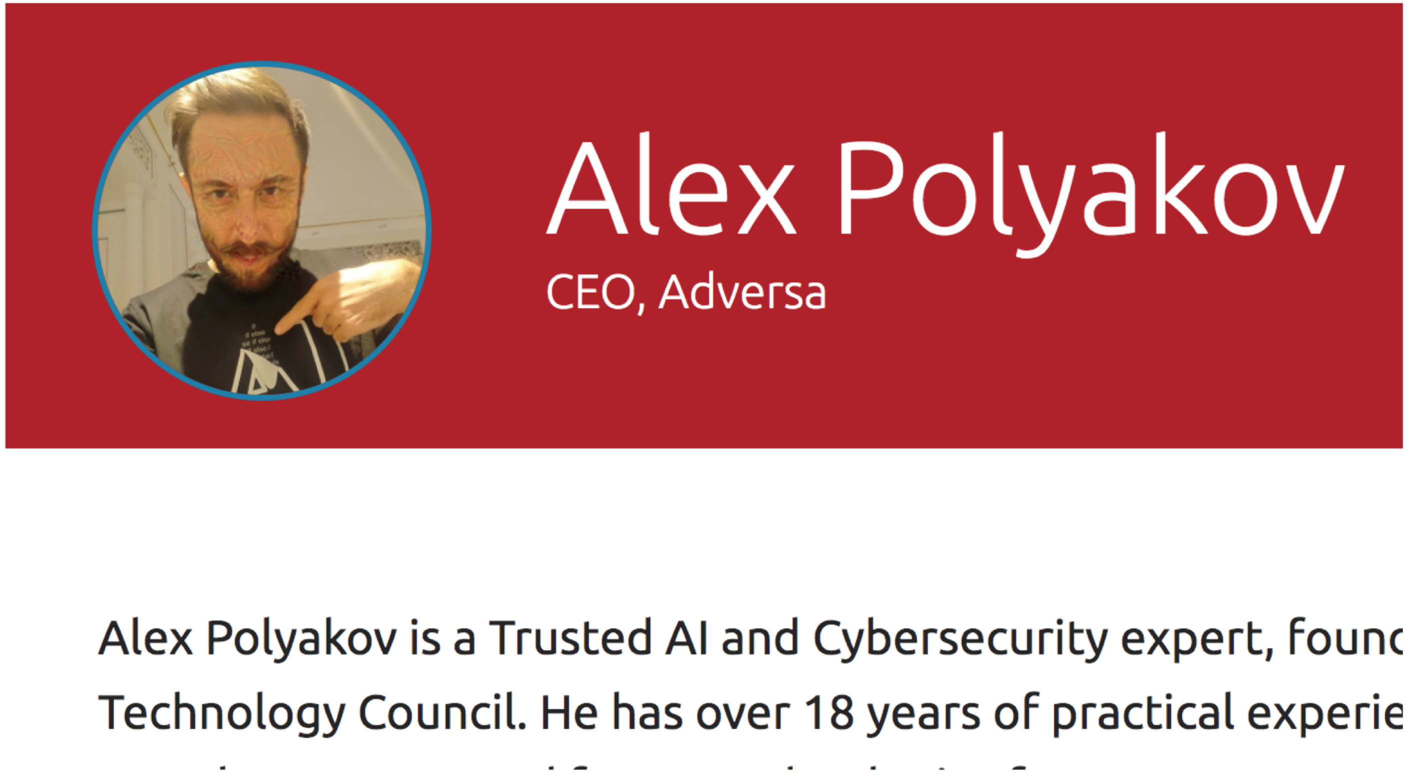
MORE LIKE THIS

Tech

NYPD's Sprawling Facial Recognition System Now Has More Than 15,000

More on this: https://adversa.ai/face-recognition-attack-adversarial-octopus/

Digital Attack Demo (Who is on this photo?)



Testing Physical Facial Recognition

ADVERSA

Story

- A physical security solution provider expressed security concerns
- They needed to find the most reliable hardware/software solution
- Real-world testing of products against **physical attacks** was required

Our Goal

To demonstrate the real threat and test as much various conditions as possible.

- Must work in physical environment
- Must be transferable as we had a blackbox threat model
- Must be imperceptible as much as possible to avoid suspicion

Existing Research

- 100+ research papers on facial recognition
- Potential attacks were published few times in media
- No practical live demos/code of physical attack

Why to Test in Real Environments

- Reality is more complicated than lab conditions
- Different camera/environment/preprocessing features
- All that should be taken into account

Real Environment Conditions

Environment features

- *Light, brightness, etc.*
- *Distance to object*

Device features

- *Resolution quality*
- *Color rendering*

Preprocessing features

- *Codecs compression*
- *Data transfer compression*

Real Environment Attack Approaches

- Fine function for big pixel difference
- Train with various sizes and angles
- Add or subtract color changes
- Gaussian blur generation
- Color randomization

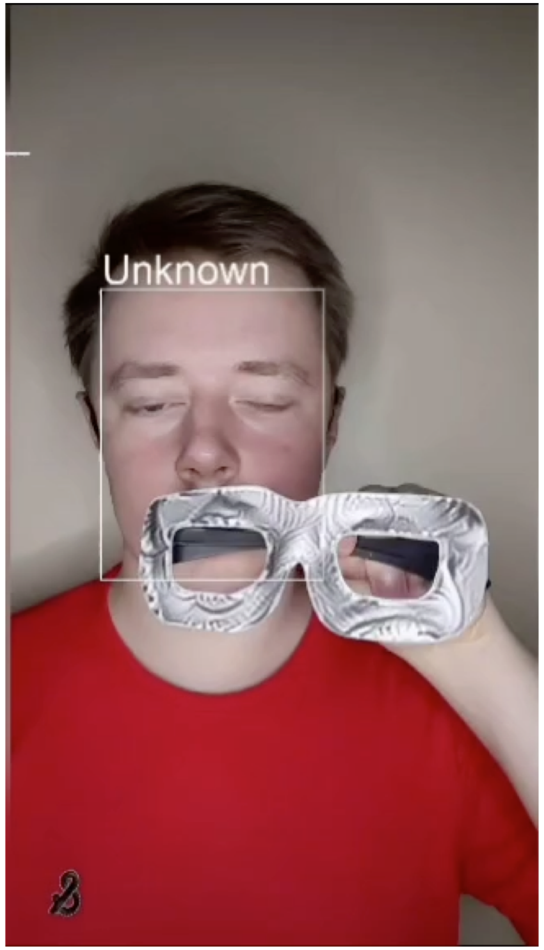
Working Combo (one of the possible)

- For better Accuracy
  - *Calculate adversarial changes for each layer of a neural network (Deepfool)*
  - *Use ensemble of networks to train attack (ResNet, SENet, FaceNet)*

- For better Transferability
  - *Random noise while constructing patch*
  - *Use various face detection frames to transfer between face detection algorithms*

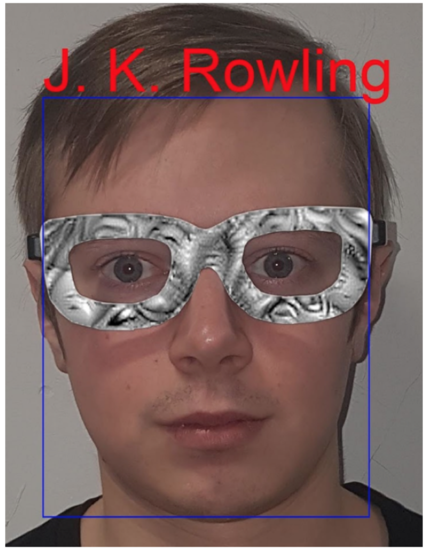
- For better Imperceptibility
  - *Use smoothing function (TVLoss)*
  - *Use black and white glasses*

Attack Demo



Attack Results

- Physical attacks have successfully fooled a facial recognition system
- Glasses have achieved the best misclassification rate
- They work even on Man to Woman and other combinations
- Further optimization is possible but was out of scope



Testing Facial Recognition: Defenses

How to protect your systems

ADVERSA

Why Test Defenses

This is not an SQL injection with a known protection way

- *You patch against these glasses and I create new ones*

No one-size-fits-all protections

- *Must understand threat model (devices/ APIs / models)*

No 100% reliable protections

- *Must test combinations and accept trade-offs*

Defense Approaches

- Modify training phase (adversarial training, distillation)
- Modify models (different activation functions, layers)
- Modify model inputs (JPEG encoding, compression)

Defense Comparison Results

- Modifying training is expensive and can be bypassed
- Modifying models may lead to decreased accuracy
- Modifying inputs is good but complicated and task-specific

Secure AI lificycle

Awareness

- Educate stakeholders in AI and security teams about AI threats
- Study relevant policies and practices for secure and trusted AI

Assessment

- Perform initial threat modeling to understand AI security risks
- Conduct initial security testing for mission-critical AI systems

Assurance

- Understand and respond to AI security findings
- Integrate security activities into AI development lifecycle

Next Steps

- Educate your AI and security teams
- Perform initial Threat Modeling
- Conduct AI Red Teaming
- Understand and address security findings
- Integrate these steps into your security lifecycle

Thank You!

Download this study: https://adversa.ai/faces

Talk about AI Security: info@adversa.ai

ADVERSA