COMMSEC

HITBSECCONF AMSTERDAM - 2021

The History of Adversarial AI

Eugene Neelou & Alex Polyakov
Adversa AI, Israel

Alex Polyakov

- CEO & Founder: Adversa.AI
- 18 years in Cybersecurity, 6 years in AI
- Member: Forbes Technology Council
- Author: 2 books, First Adversarial ML MOOC
- Speaker: 100+ conferences in 30+ countries
- Interests: Trusted AI, SynthBio, Neuroscience

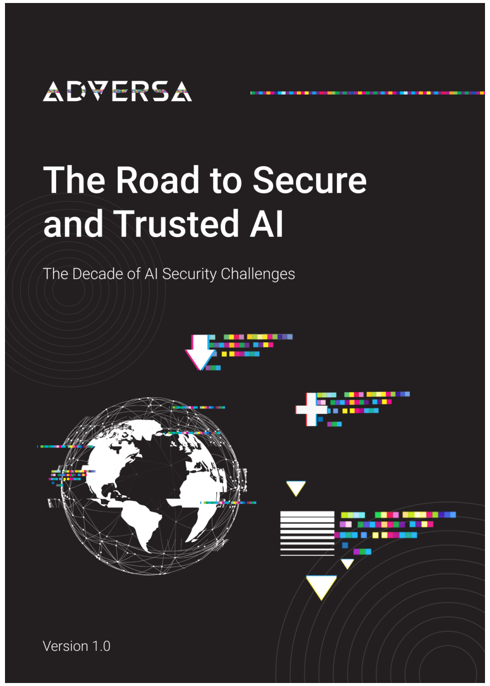
Eugene Neelou

- CTO & Co-Founder: Adversa.AI
- 13 years in Cybersecurity, 6 years in AI
- Ex-Director of Security Research & Data Science
- Product leader, security researcher, consultant
- Completed 100+ projects in 30+ countries
- Expert in turning research into products

We are on a mission to increase trust in AI by protecting it from cyber threats

Agenda

- The Need For Secure AI
- The Progress Toward Secure AI
- AI Systems Under Attacks
- How To Attack AI Systems
- *Case Study: Security Of AI Facial Recognition*
- How To Defend AI Systems



Report: https://adversa.ai/hitb

The Need For Secure AI

- The Next Decade Of AI Security
- AI Is The New Attack Vector
- Real Incidents In AI Systems
- The Inception Of Trustworthy AI

The Next Decade of AI Security

- 1990s – Network security
- 2000s – Endpoint security
- 2010s – Application security
- 2020s – AI security (Software 2.0)

AI Is The New Attack Vector

Traditional Software

- **Powered by**
  Fixed program logic
- **Workflow**
  Tasks and commands
- **Interaction**
  Graphical UI with menus and buttons
- **Typical problems**
  Improper validation, access control issues, system & security misconfiguration

AI (Software 2.0)

- **Powered by**
  Flexible ML training
- **Workflow**
  Learning and decisions
- **Interaction**
  Cognitive UI with visual, audio, text commands
- **Typical problems**
  Model manipulation, data exfiltration, model & data infection

Real Incidents In AI Systems

Confidentiality

- Personal data extraction from Netflix statistics during ML contest
- Model cloning of pre-released GPT-2 model from Open AI

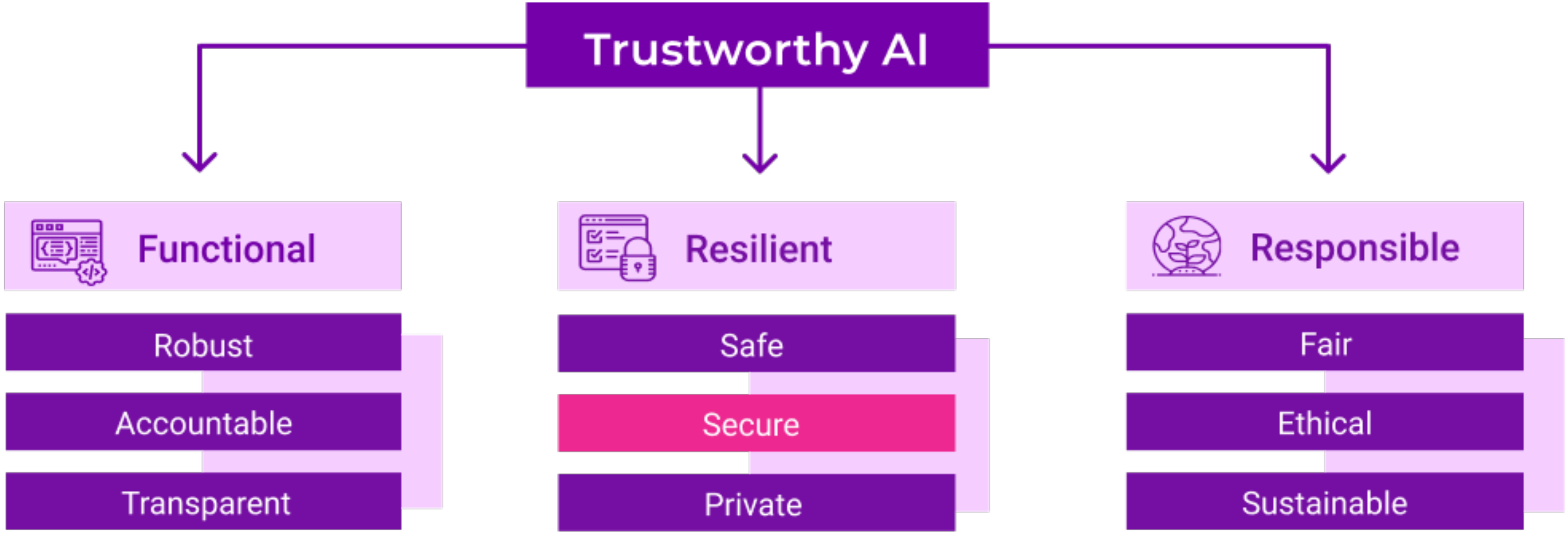
Integrity

- Evasion of Cylance AI-based malware detection
- Poisoning of VirusTotal dataset with fake samples

Availability

- Fooling Tesla autopilot could lead to self-driving car crash
- Denial of service in IoT via resource exhaustion attacks

The Inception Of Trustworthy AI

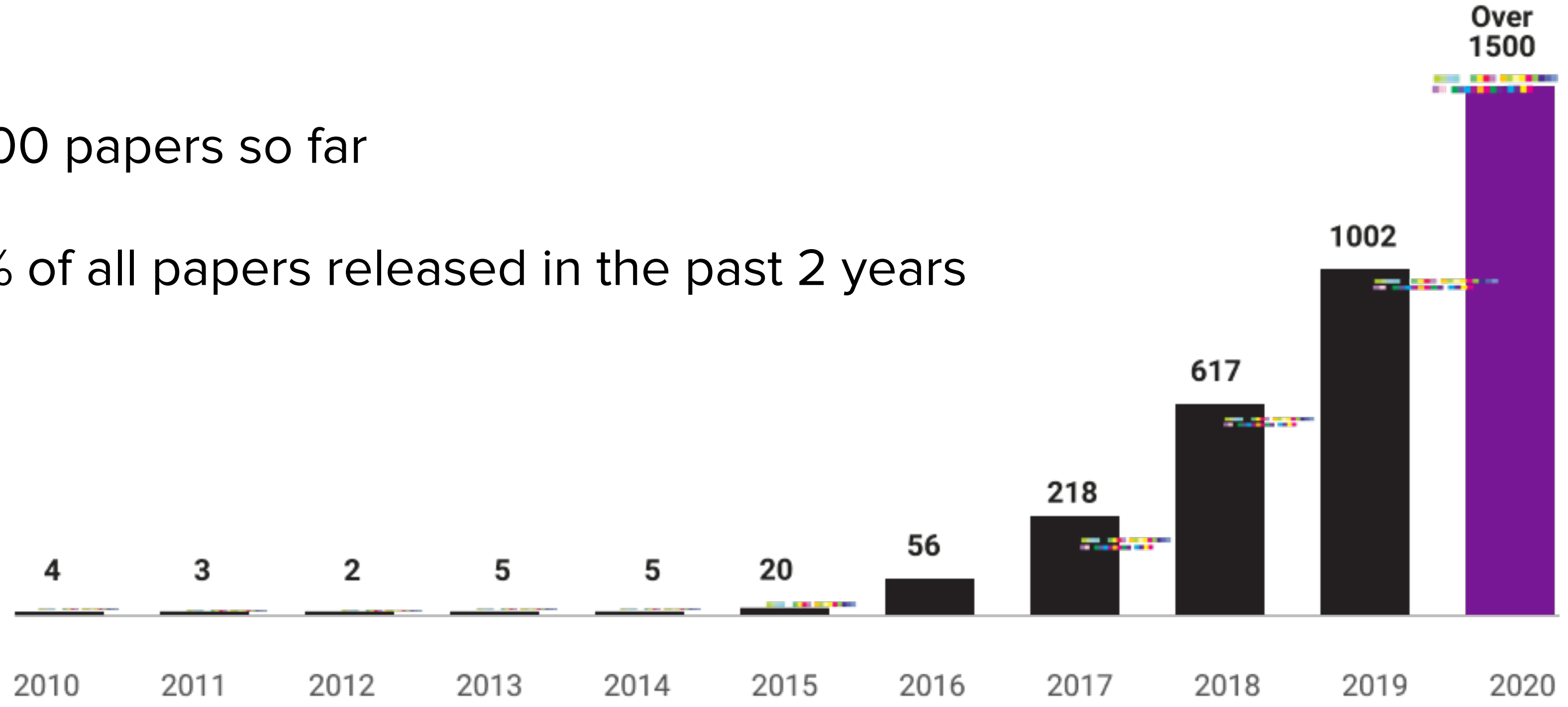


The Progress Toward Secure AI

- Adversarial Machine Learning Research
- Most Active Country Contributors
- Key Events In Adversarial ML

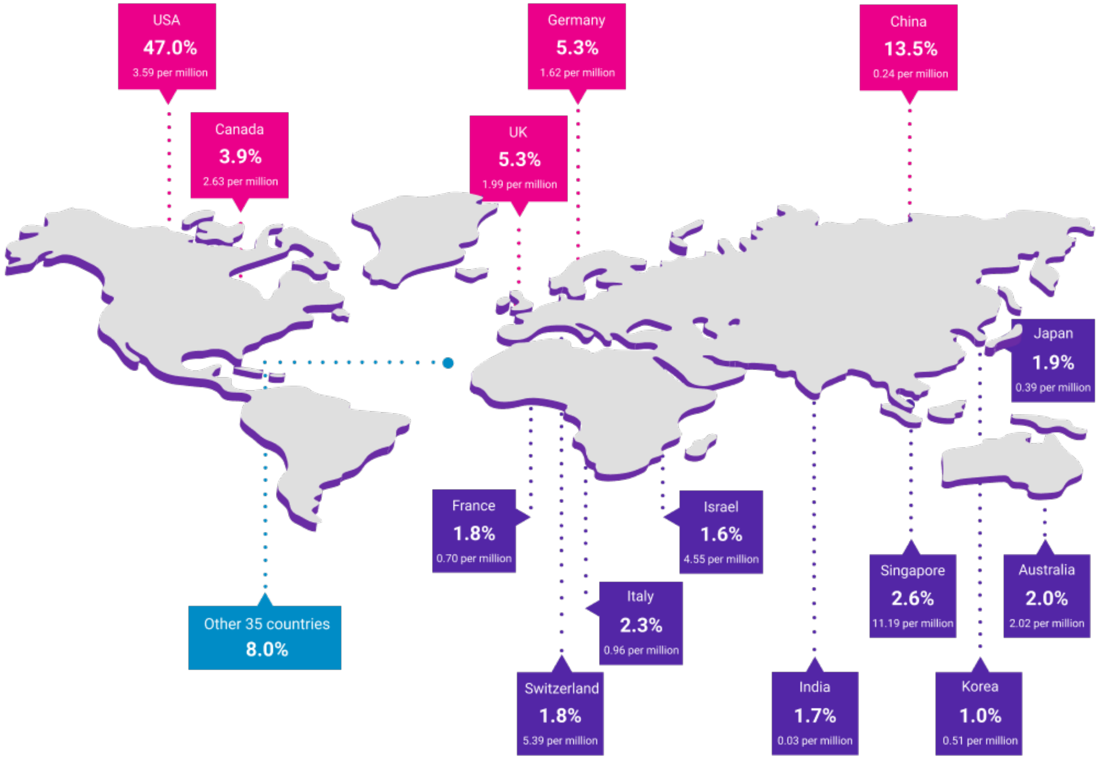
Adversarial Machine Learning Research

- Over 4,000 papers so far
- Over 50% of all papers released in the past 2 years



Most Active Country Contributors

- Contributors from 50 countries
- Over 90% of papers from 14 countries



Key Events In Adversarial ML

- **Academia**

  2004 – early works in adversarial machine learning

  2014 – reborn interest with attacking deep learning

  2021 – more than 4000 research papers released

- **Governments**

  2016 – first national AI document mentioning security in US

  2019 – first national AI document focused on security in US

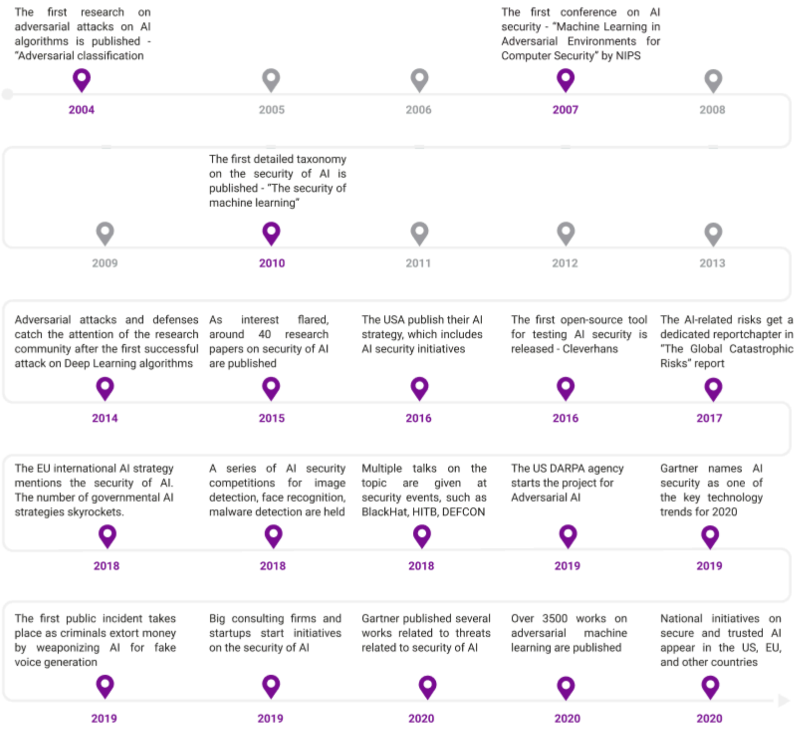
  2021 – first international regulation for high-risk AI in Europe

- **Industry**

  2018 – First trusted AI consulting and startups

  2019 – Gartner starts covering the AI security topic

  2020 – Growth of AI red teams (FB, MS, Nvidia, Open AI, MITRE)



AI Systems Under Attacks

- AI Areas Under Attacks
- AI Datasets Under Attacks
- AI Applications Under Attacks
- AI Industries Under Attacks

AI Areas Under Attacks

Vision 65%

Image classification, Object detection, Medical imaging
Video analytics, Action recognition, Person re-identification
Facial recognition, Biometric analysis (emotion, gender, age)

Analytics 18%

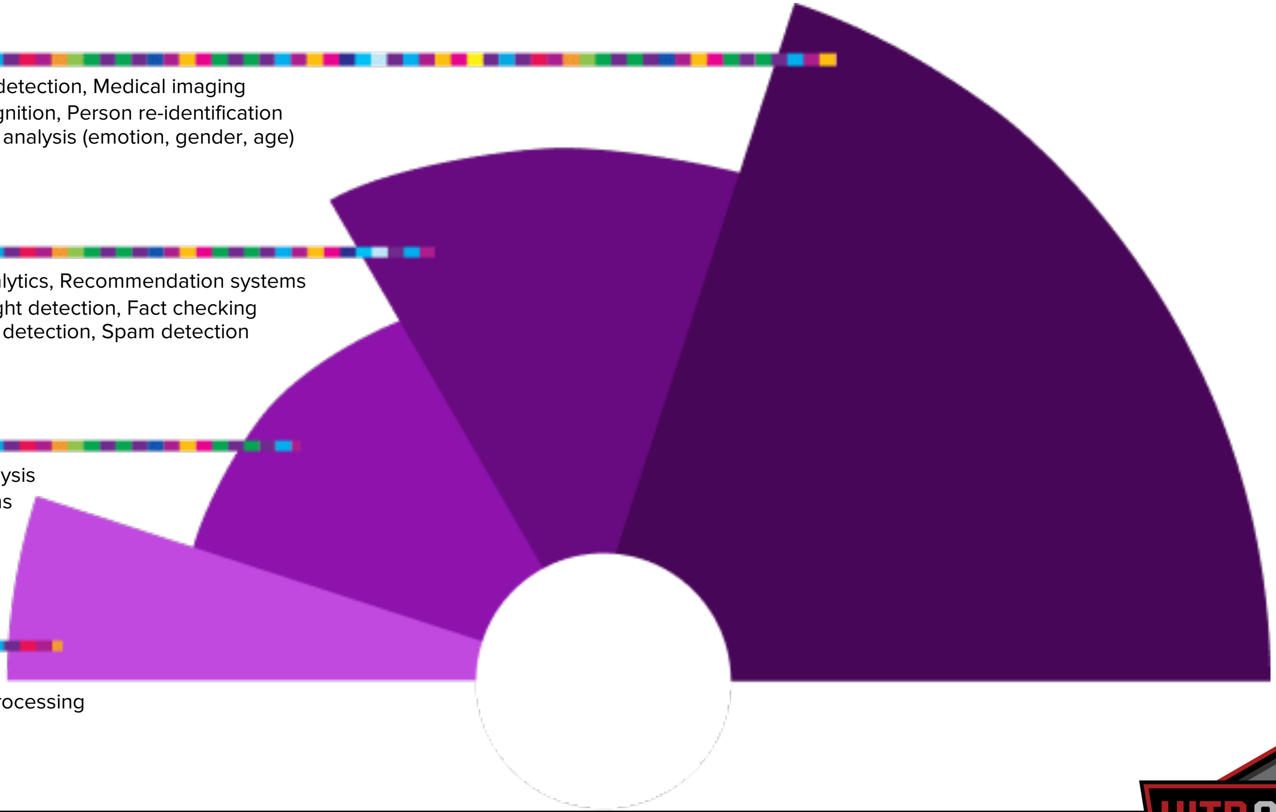
Records analytics, Audio analytics, Recommendation systems
Content moderation, Copyright detection, Fact checking
Malware detection, Intrusion detection, Spam detection

Language 13%

Speech recognition and analysis
Text analytics, Dialog systems

Autonomy 4%

Reinforcement learning
Sensor perception, Signal processing



AI Datasets Under Attacks

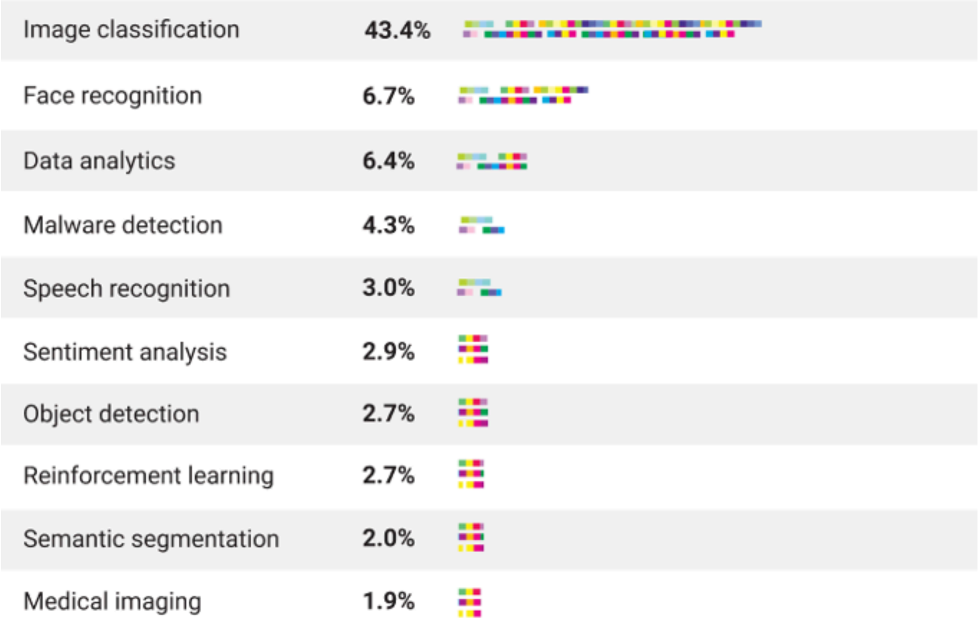
- The dominance of attacks against AI systems with image processing shouldn't mislead you into thinking that other AI applications are less vulnerable.

- Counterintuitively, AI applications with fewer attacks might be at greater risk because the interest to develop defenses for them is significantly smaller.

Image	60.8%
Text	10.0%
Record	5.5%
Binary	4.3%
Audio	4.1%
Graph	3.2%
Signal	3.0%
Agent	2.8%
3D	2.2%
Traffic	2.1%
Video	1.9%

AI Applications Under Attacks

- Over 2000 successful attack cases against over 100 unique AI applications
- AI applications were *never* inherently robust
- **Attacks are transferable across applications**



AI Industries Under Attacks

- Attack Research shows original interest of researchers to attack AI in a given industry
- Transferable Risk reveals the real threat landscape based on our risk correlation
- **Attacks are transferable across industries**

Position	Industry	Attack Research	Transferable Risk
1	Internet	23%	97%
2	Cybersecurity	17%	41%
3	Biometrics	16%	67%
4	Automotive	13%	79%
5	Healthcare	9%	87%
6	Industrial	5%	74%
7	Smart Home	5%	89%
8	Retail	4%	86%
9	Finance	4%	95%
10	Surveillance	3%	77%
11	Robotics	1%	61%

How To Attack AI Systems

- Focus Of Adversarial ML Research
- Categories Of AI Attacks
- Adversa Top 10 Attacks

HITB SECCONF
AMSTERDAM - 2021

Focus Of Adversarial ML Research

- **Attacks (49%)**

Researchers invent new ways of hacking AI models, AI infrastructure, side channels, and software bugs.

- **Defenses (47%)**

Researchers suggest defenses like analysis of data inputs/outputs, algorithm modifications, or retraining.

- **Surveys (3%)**

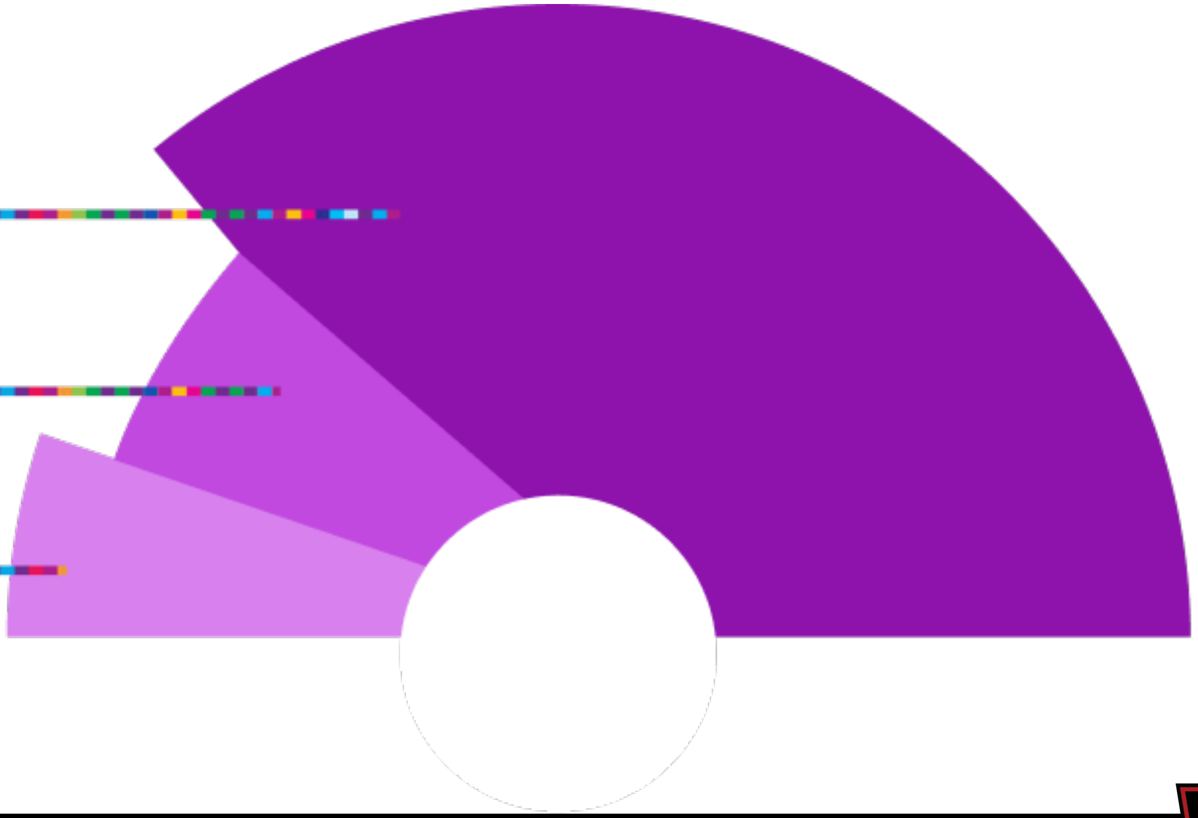
Researchers review security threats for AI systems and investigate existing attacks and defenses.

- **Tools (1%)**

Researchers develop tools for performing vulnerability testing and verification of AI models.

Categories Of AI Attacks

Manipulation	82.6%	Evasion attacks, Reprogramming
Infection	9.5%	Poisoning, Backdoors, Trojans
Exfiltration	7.8%	Inference attacks, Model stealing



Adversa 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

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 systems

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 threat actors to repurpose AI models and make them execute unexpected tasks

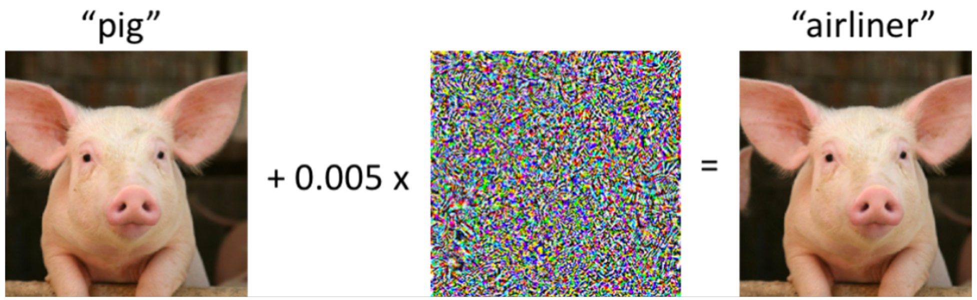
HITB SECCONF
AMSTERDAM - 2021

Case Study: Security Of AI Facial Recognition

- Intro To Evasion Attacks In Computer Vision
- Problem Of Securing AI Facial Recognition
- AI Security Testing Challenges
- AI Security Testing Results

Intro To 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
- Model makes a wrong prediction controlled by an attacker
- Changes are imperceptible for system owners



Case Study Profile

- A smart home solution provider expressed security concerns
- They needed to implement a secure facial recognition solution
- The goal was to find the most reliable hardware/software vendor

Problem Of Securing AI Facial Recognition

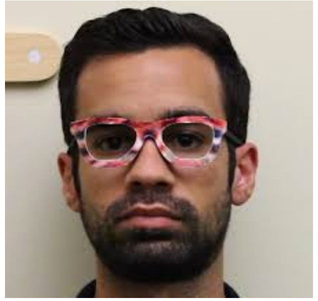
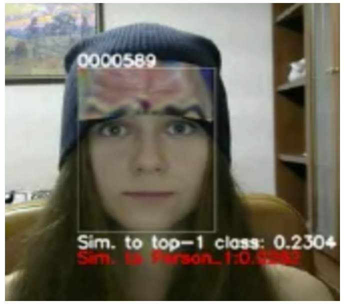
- Over 4000 research papers on adversarial machine learning
- Over 100 research papers on facial recognition security
- Countless combos of different conditions (attacks, models, environments)
- No clear understanding of real-world risks beyond research

AI Security Testing Challenges

- **Goals**: Confidence reduction, untargeted/targeted misclassification, etc.
- **Forms**: Glasses, lenses, mask, hat, moustache, band-aid, etc.
- **Knowledge**: Black-box, gray-box, white-box
- **Constraints**: Time, computation resources, data
- **Conditions**: Printing quality, inconsistent colors, position, angle, size, etc.
- **Algorithms**: FGSM, BIM, PGD, EOT, etc.

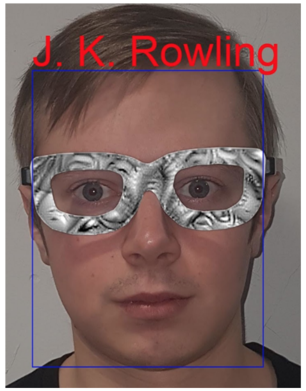




AI Security Testing Results

- Physical attacks have successfully fooled facial recognition systems
- Glasses and bandanas have achieved the best misclassification rate
- Further attack optimization was possible but was out of scope



How To Defend AI Systems

- Lifecycle for Secure AI Development
- Your Next Steps

HITB SECCONF AMSTERDAM - 2021

Lifecycle for Secure AI Development

1. Identify

Understand current AI security posture with *asset management*, *threat modeling*, *risk assessment.*

2. Protect

Implement protective controls such as *security awareness*, *system hardening*, *secure AI development.*

3. Detect

Defend against active adversaries with *security monitoring, threat detection*, and *AI red teaming*.

4. Respond

Prepare for AI security incidents by developing practices of *AI incident forensics and response.*

Get detailed lifecycle: https://adversa.ai/hitb

Your Next Steps

1. Awareness

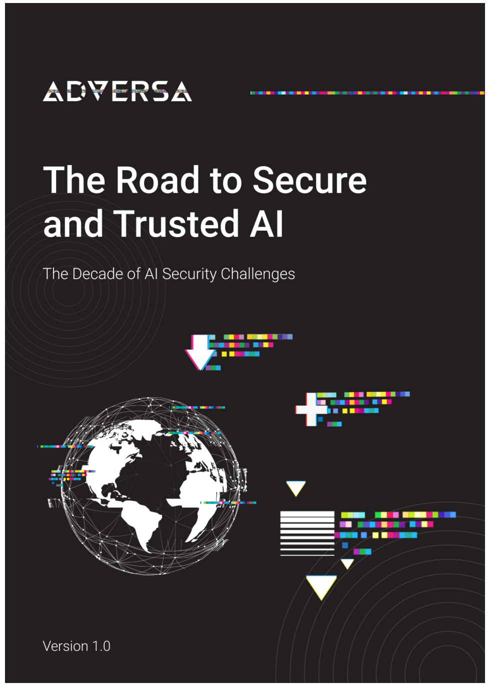
- Educate stakeholders in AI and security teams about AI threats
- Study relevant policies and practices for secure and trusted AI

2. Assessment

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

3. Assurance

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



Report: https://adversa.ai/hitb

Thank You

Follow us! adversa.ai/stay-updated

Collaborate with us! info@adversa.ai