ADVERSA

The Road to Secure and Trusted AI

The Decade of AI Security Challenges

Version 1.1

Table Of Contents

Introduction

Why Security Of AI Matters

What AI Areas Are In Danger

Who Is In Danger Exactly

When AI Security Started

Where AI Security Risks Are Addressed

How AI Attacks And Defenses Are Launched

The Demand For Secure AI

Since the inception of the internet, every decade has been dominated by a specific technology paradigm. Inevitably, cyber attacks and hackers have evolved to exploit these new technologies, with new defenses and security solutions being developed as a response. While the 1990's were a decade of network security, the 2000's were dedicated to endpoint security, and the 2010's were focused on securing applications. The scientific and commercial interest is shifting from regular software to artificial intelligence. Undoubtedly, the 2020's will be all about AI adoption followed by inevitable attacks and defenses for AI systems.

Alongside the hype around AI, it has proven its value in real-world practical applications in various areas of life and business. AI use case opportunities expand with the help of smart assistants processing speech, text, video, behavioral and other types of data. AI has introduced a new form of interface and interaction between the digital and the physical realms. People can communicate with AI-based cognitive interfaces instead of classic visual interfaces with menus and buttons.

In the old paradigm, software vulnerabilities often occurred due to improper filtering of commands, incorrect data handling, or design flaws. But now, AI commands can be visual, audial, textual or environmental. This makes it more difficult to filter, handle, and detect malicious inputs and interactions. Unfortunately, the AI industry hasn't even begun to solve these challenges yet, jeopardizing the security of already deployed and future AI systems.

The AI threat landscape will continue to develop, and it will evolve at machine speed. Cyber criminals will use new techniques to tamper with mission-critical AI solutions, whether in self-driving cars, facial recognition systems, autonomous drones, robots, smart assistants, or financial algorithms. Eventually, threat actors will weaponize AI for malicious purposes.

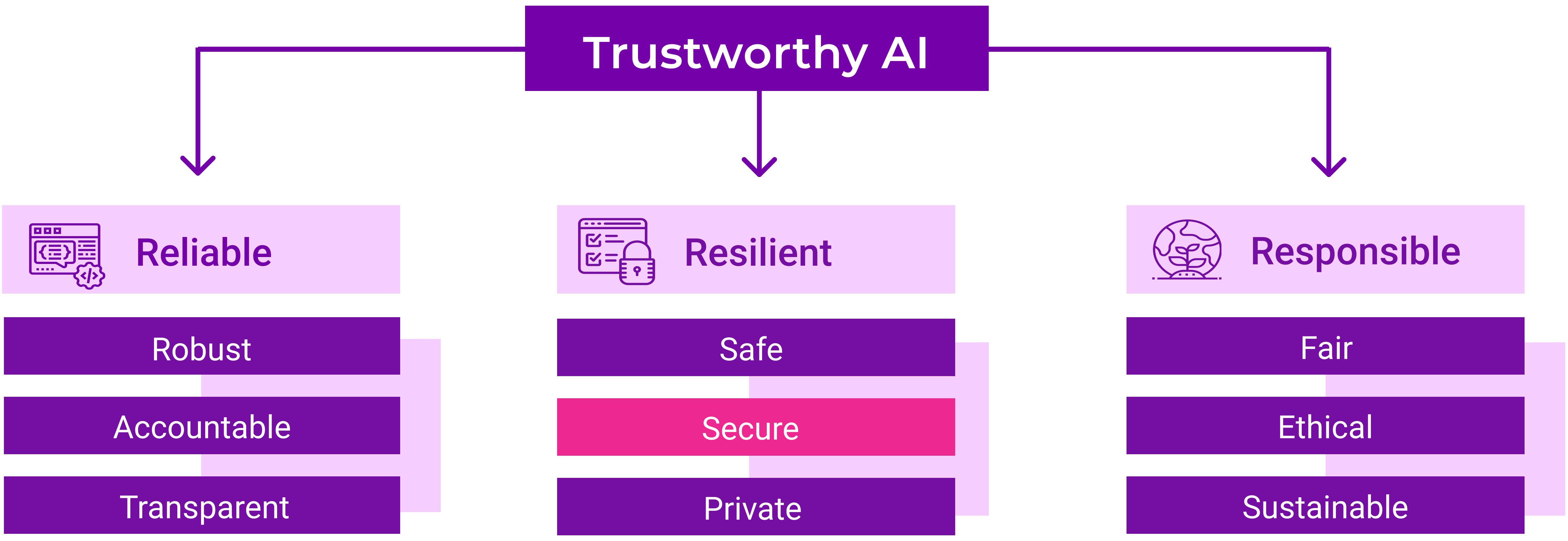
As a team of experts from a Trusted AI research and advisory company, we have conducted detailed research and discovered that recent initiatives from governments, academia, and industry have highlighted the immediate need for the security of AI.

This report sheds light on the field of AI security and adversarial ML of the previous decade from 2010 until 2020. We want to make companies aware of cybersecurity problems in AI systems they develop or use by analyzing this area from different angles. Our findings demonstrate the urgent need for attention to the AI security problem and terrifying unpreparedness of AI innovators to develop a safe, secure, and trusted AI.

The Inception Of Trustworthy AI

Building trustworthy AI is essential for growth and wide adoption of smart technologies by organizations and societies. They should provide and use AI solutions that not only perform the task but also do it in a reliable and ethical way. It is crucial to address the problem of trusted AI systematically from different angles and perspectives.

We've collected 9 characteristics that AI should have based on governmental, business, and academic works. We've grouped them into 3 groups for easy memorization.



Trustworthy AI

Reliable: Robust, Accountable, Transparent

Resilient: Safe, Secure, Private

Responsible: Fair, Ethical, Sustainable

- **Robustness** means abilities of AI systems to function as intended, behave predictably especially in the situations of uncertainty, and fail gracefully in case of fatal errors.
- **Accountability** explains how to operationalize AI systems in a responsible way by providing human oversight and control over AI behaviors and outcomes.
- **Transparency** introduces mechanisms for traceable and explainable AI decisions as well as policies that respect human rights to know when they interact with AI systems.

- **Safety** is required for the mission-critical situations and human-computer interactions, where an AI system should prevent any harm to living beings or the environment.
- **Security** makes AI systems resilient to malicious activities and adversarial attacks such as general software security attacks or specific attacks against AI algorithms.
- **Privacy** demands AI systems to implement adequate data governance mechanisms, data protection and access controls for collected and inferred data.

- **Fairness** enables inclusion, diversity, and accessibility by introducing practices for examining bias and counting interests of everyone affected by AI systems.
- **Ethics** touches on moral principles of interaction between humans and AI systems, which should respect human rights and empower people, not the opposite.
- **Sustainability** lets organizations use AI algorithms to make people's lives better by meeting today's needs without compromising the interests of future generations.

All of these requirements are equally important for building beneficial and trustworthy AI. In this report, we focus on the AI security domain called adversarial machine learning.

AI Industry Voice

Ariel Herbert-Voss

Research scientist at OpenAI, CS PhD at Harvard, co-founder of Defcon AI Village

Avivah Litan

Vice President and Distinguished Analyst in Gartner Research

Battista Biggio

Assistant Professor at University of Cagliari, Co-Founder of Pluribus One

Davi Ottenheimer

Trust & Digital Ethics Technologist, Advisory board member of Accenture, ex-Director of Trust in EMC

Nikita Lukianets

Founder at Open Ethics Initiative for AI transparency, Founder & CTO at PocketConfidant AI

Martin Szugat

Program Chair of conferences Deep Learning World and Predictive Analytics World

Nigel Willson

Speaker, AI ethics advisor, founder of AwakenAI.org, co-founder of WeandAI.org

Oliver Rochford

Researcher, Strategy advisor, former Gartner Analyst

Roman V. Yampolskiy

CS Professor, AI safety & security researcher, Author of the book "Artificial Superintelligence: a Futuristic Approach"

Vered Shwartz

Researcher at Allen Institute for AI and University of Washington

WHY should companies care about responsible AI and especially the security of AI systems?

If AI is a core part of your product, manages a portion of your financial system or business strategy, it automatically becomes a target by anyone looking to make a shady buck off your company. Being able to keep your AI systems safe helps protect your company and your customers - just like investing in good security tools, practices, and people for the rest of your tech stack.

Threats against AI are not new, but have been insufficiently addressed by enterprise users. Malicious hackers have attacked AI systems for as long as AI has existed. The same can be said for benign actors who introduce mistakes and biases that undermine model performance and fairness.

The reason is twofold. It is an investment to mitigate a potential security risk, as tomorrow AI may become the weakest link in the security chain and be exploitable by attackers with weaponized, scalable attacks... The second reason is that governments may require AI to be trustworthy via specific regulations and standards.

Security and trust are an imperative for artificial intelligence, there is already much reported in the press on the negatives side of AI which has led to lower than desired adoption. When implemented badly AI can cause huge reputational risk and damage...

AI systems are software systems, without appropriate levels of security they can't function and deliver benefits to the users, in fact they can become quite harmful to the users, company's reputation and stockholder value.

Building trust in the security and safety of machine learning is crucial. We are asking people to put their faith in what is essentially a black box, and for the AI revolution to succeed, we must build trust. And we can't bolt security on this time. We won't have many chances at getting it right. The risks are too high - but so are the benefits.

The two main drivers will be governmental regulation, especially in already highly regulated industries such as pharma, insurance or finance, and user acceptance.

WHAT should motivate companies to invest in AI security initiatives?

Already there is a budget allocated to security of infrastructure and I would see it as a carve-out. It'll be more of a question of how to justify a shift in resources, rather than a whole new allocation; for example, spend less on anti-virus and move that line item to AI.

AI Industry Voice

WHAT should motivate companies to invest in AI security initiatives?

If they retain that securing AI can be a valuable asset for them. This depends clearly on what they use AI for, and if attacking AI may be a realistic threat for their businesses.

Secure AI systems are part of Responsible AI (or Ethical AI) systems and can comprise of Trust, Transparency, Fairness, Security and Accountability. These should all be at the centre of any AI solution or initiative primarily because a company or organisations culture should want to promote these things, not because they are waiting for legislation to be told they have to.

Make sure you invest in initiatives that fit a realistic threat model - not all AI systems are equal targets. Depending on the system, the target might be the training data containing sensitive customer information or the model weights that enable you to make unique predictions that form the core of your product.

Large enterprises with a strong governance already invest in AI security. Other companies will follow when the incidents reported in the media are increasing.

AI Security is not some bonus feature, it is a fundamental requirement for every AI-based project. Without security, quality AI is not possible, which harms both reputation and profit potential.

HOW should companies make their first steps toward secure and trusted AI?

The first steps always are awareness, recognizing there is a problem opens the door to addressing it with a solution. The most likely person to buy a security solution is the one who says they refuse to hear about a problem unless there is a solution ready.

Start by making sure the data you use to train your AI systems is traceable. The most exploited attack vector for AI systems is the data processing pipeline. Steer clear of buying products that rely on FUD marketing and focus on identifying real threat potential.

Leaders should invest efforts to bring the team together, and to openly talk about risks that users are facing when using the product...Besides the organizational efforts, there are structural efforts that could be done to make products more transparent to the end users, such as through self-disclosure.

...Thorough and robust testing is also important, in the current fast moving world there is a desire to deploy fast and fix later, with AI this can easily put your project or organisation at risk, and even criminal prosecution.

First, AI security design, checks and audits must be an essential part of the AI product life cycle. Second, in addition to data governance organizations must establish AI governance processes, guidelines etc. Third, companies must hire (internal or external) security experts as AI validators.

Consider a threat-model analysis of their business where AI is involved. Evaluate the corresponding risks, identify and prioritize which AI components to secure. Use mitigations. In the end, this is equivalent to “pentesting AI” and one may not have in-house know-how and people for that. Hence, the company may need a consultancy service or anyway rely upon an external service to automate the process within their development pipeline.

...Don’t wait until the inevitable breach, compromise or mistake damages or undermines your company’s business, reputation or performance. Secure your AI today.

ML-based products need to be tested for correctness, robustness, fairness, security, and privacy aspects before they are deployed. The pace of research in fields that use ML (NLP, vision) has accelerated, and it's largely a positive thing. But I think it's beneficial not to rush to deploy every new technology. The lack of transparency of neural models makes them vulnerable to various types of attacks we might not yet be aware of.

Executive Summary

1

The future is now

AI has disrupted many industries and is the core of next-generation technologies. Like any maturing technology, AI is becoming a lucrative target for cybercriminals.

2

Skyrocketing AI incidents

The number of real-world AI incidents is growing in automotive, biometrics, robotics, and internet industries. Problems in confidentiality, integrity, and availability prove the need for secure and responsible AI.

3

AI applications under attack

The most targeted AI area is computer vision followed by analytics and language. This is backed by the most scrutinized applications such as image classification, facial recognition, malware detection, speech recognition, and data analytics.

4

AI-powered industries at risk

Being early AI adopters, the most scrutinized industries are internet, cybersecurity, and biometrics. Yet, AI attacks are transferable across industries, expanding the attackers' attention to automotive, smart home devices, and finance industries.

5

AI security research on the rise

The security of AI remained a niche research field for over a decade. The recent exponential growth of AI has motivated governments, academia, and industry to publish more research in the past 2 years than in the previous 2 decades.

6

International AI contributions

The USA was the first to introduce a national AI strategy with dozens of countries following their path. This has also driven the R&D field, with the AI arms race between USA and China publishing more research on the security of AI than the rest of the world combined.

7

Vulnerabilities in AI systems

Most attacks are aimed at manipulating behavior of AI systems. This is followed by interest in the AI model's internals and data exfiltration as well as the infection of AI models and datasets. Top 10 attack methods represent each of these categories.

8

Lifecycle for AI security

To protect AI from attacks, a comprehensive security program is required. Solutions are required for security testing, hardening, detection and response stages. Unfortunately, the number of attack techniques is bigger than defensive ones, and defenses are not evenly distributed across cybersecurity lifecycle stages.

9

Action required today

The AI industry is terrifyingly unprepared for real-world attacks against AI systems. AI is essential to global strategy but AI revolution cannot succeed without trust. Public perception on trustworthy AI will be a core driver determining whether societies and businesses will adopt AI for good or face another AI winter.

Progress Toward Secure AI

Here we highlight the progress toward secure AI across industry, government, and academia.

	Past	Present	Future
Industry	Little knowledge. Absent, manual or ad-hoc testing of corner cases and attacks, few basic open-source tools for automating attacks and defenses.	Gartner highlighted AI security as a strategic trend for 2020. Most auditing firms such as Deloitte, EY, KPMG, PwC, Accenture, McKinsey started offering consulting for secure and trusted AI.	New startups will appear and focus on security of AI. We will see the market segmentation based on parts of AI ecosystem. Market size for the security of AI is expected to be $2-6B by 2025.
Government	US, EU, and China start national AI initiatives that feature aspects of AI security and AI trustworthiness.	Governments include secure AI in their national AI strategies. Leaders launch more focused initiatives: DARPA AI security grants and GARD program, EU guidelines for trusted AI.	Number of security-focused national AI policies and regulations will grow significantly in the next years as a follow-up to countries' initial AI strategies.
Academia	Before 2019, in 10 years less than 1000 research papers were published. Works are exploratory and mostly on toy models and datasets.	In 2019 alone, over 1000 research papers were published with hundreds of attacks on real-world applications from FAANG, BAT, and top CS universities.	With the growth of the AI security field, adversarial attacks will expand the attack surface to the whole AI ecosystem. The total amount of research will double in coming years.

AI Security Incidents

The elemental parts of every system's cybersecurity are confidentiality, integrity, and availability. They define desired features of systems that handle data properly while restricting unauthorized access, ensure reliable behaviors and outcomes, and keep systems and data operational and accessible at all times.

These core principles are relevant for AI-powered systems as the incidents below show.

Confidentiality

Netflix faces a privacy disaster

This incident took place back in 2006 when the information about movies rented by Netflix subscribers was shared for research but third parties were able to recover user details including watching history, personal preferences, and behaviors.

Target knows you better

Target's intrusive advertisers turned a purchase history into pregnancy prediction algorithm to send coupons to future mothers. It unpleasantly confused families as the organization knew about their child-bearing in advance.

Integrity

Facial recognition is subverted by protesters

In Hong Kong, protesters have grown increasingly concerned that police abuse facial recognition software to make arrests. To avoid detection, many of them used scarfs, masks, 3D-printed glasses, and even hairstyling and makeup as a disguise.

Microsoft's AI chatbot learns racism from Twitter

Pretty soon after Microsoft launched a Twitter bot Tay, people started tweeting the bot with all sorts of misogynistic, racist, and Donald Trumpist remarks. Tay assimilated the internet's worst tendencies into its personality.

Availability

Autopilot keeps crashing Tesla cars

2016 was notorious for Tesla due to their car crash when the car didn't recognize a van stopped in the lane as an obstacle. There were also two other similar Tesla's autopilot accidents in 2016 and 2019 resulting in human deaths.

LG robot Cloi fails publicly

Cloi, LG's smart home assistant, was supposed to demonstrate the use of kitchen appliances. Instead, it became unresponsive and left the LG executives red-faced.

AI Areas Under Attacks

The purpose of AI is to replicate human perceptions and cognitive abilities. To teach machines to understand and act, AI researchers invented and democratized algorithms for processing images, video, audio, text, and other types of data. This has introduced a new paradigm that replaced traditional menus and buttons with cognitive user interfaces.

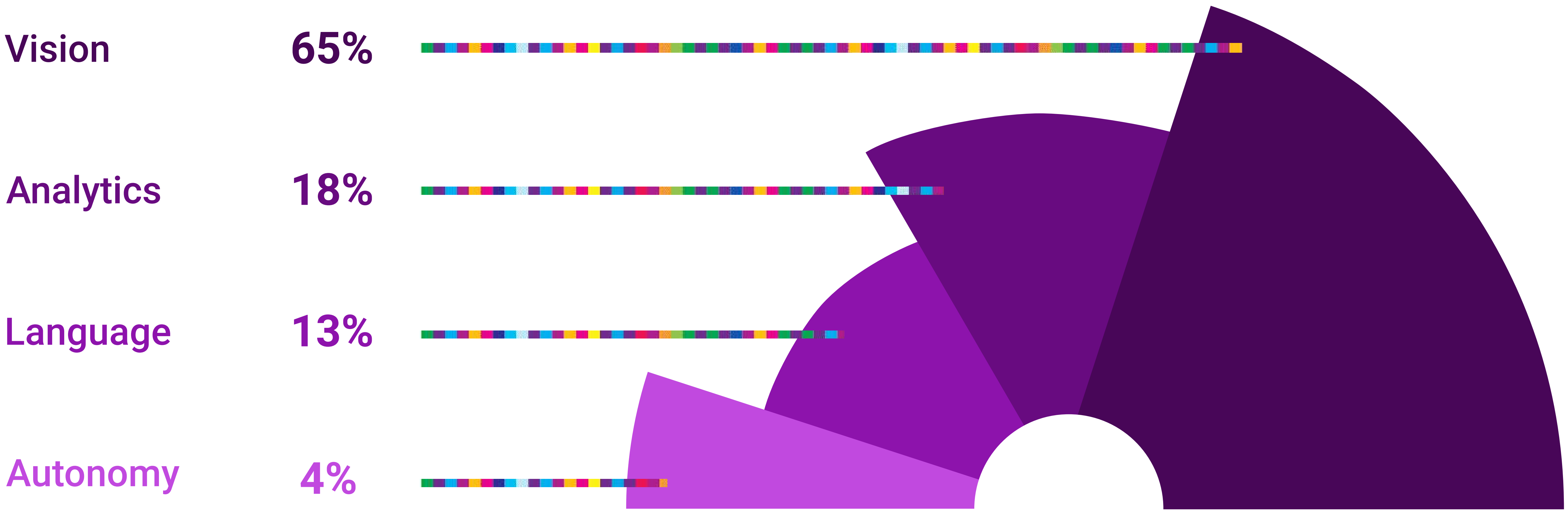
With the growth of AI, cyberattacks will focus on fooling new visual and conversational interfaces. Additionally, as AI systems rely on their own learning and decision making, cybercriminals will shift their attention from traditional software workflows to algorithms powering analytical and autonomy capabilities of AI systems.

The study of attacks and defenses for AI systems is called adversarial machine learning.

We have compiled what we believe is a complete list of close to 2000 research papers in adversarial machine learning published on the arXiv.org repository over the past 10 years. See the *Methodology* section to understand the approach, scope, and references.

The analysis has revealed the most attractive areas for AI researchers to attack.

Attacked AI areas



Takeaways

- Computer vision is the most popular AI victim with dozens of image and video applications attacked. Such interest is correlated to the maturity and popularity of the vision AI.
- While all areas of AI have been attacked already, languages and autonomy are less explored. As these domains mature, they will become attractive attack targets over time.

AI Datasets Under Attacks

Data is a vital part of every AI system, thus it should be taken into account to understand a threat landscape. Adversarial scenarios may vary from infecting data used for AI training to crafting malicious inputs interactively during AI model inference.

Image data is the most popular target because it is easier to attack and more convincing to demonstrate vulnerabilities in AI systems with visible evidence. This is also correlated to the attractiveness of attacking computer vision systems due to their rising adoption.

Yet, all other existing data types used in AI applications have been attacked already as well.

Attacked AI datasets	Share
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%

Takeaways

- 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 that experienced fewer attacks might be at greater risk because the interest to develop defenses for them is significantly smaller.

AI Applications Under Attacks

Many AI applications such as facial and speech recognition have become a natural part of our daily lives. Other use cases such as medical imaging or malware detection have disrupted their industries. Popular software has always been an expected target for cybercriminals. Certainly, AI systems are not going to be an exception.

There are over 1000 attack cases against over 58 applications in research literature published in the previous decade (see *Methodology*).

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%***

Takeaways

- The most popular AI applications are also the most scrutinized. Some such as image classification span across industries putting them at risk. Others like facial recognition and malware detection undermine trust in AI being at the core of mission-critical products.
- There is hardly an area of smart software that isn't affected. As the AI adoption grows, researchers explore cases from machine translation, lending decisions, or electrodiagnosis to emotion detection, content moderation, or fact checking in order to subvert them.

AI Industries Under Attacks

It is hard to find an industry which is not disrupted by the AI revolution. Automotive, biometrics, healthcare, internet, to name a few, are all inevitably changing the way their businesses work. They gain insights from data at scale, improve decision making, or even replace employees with AI systems that only humans were able to perform before.

Yet, just like with any other computer system, it's only a matter of time when attacks on AI systems will become as common as compromised accounts and data breaches.

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%

Takeaways

- **The Attack Research** shows the interest of researchers to attack AI applications in a given industry. It comes as no surprise that the most popular targets are the early AI adopters. They have developed mature AI systems by now and accumulated a lot of data and intellectual property, making them an attractive target for attackers.

- **The Transferable Risk** reveals the real threat landscape based on our expert matching. In fact, attacks on similar AI models across industries are comparable. For instance, attacks on facial recognition systems are transferable between biometrics and surveillance. Thus, biometric attacks may be counted toward the surveillance industry and vice versa.

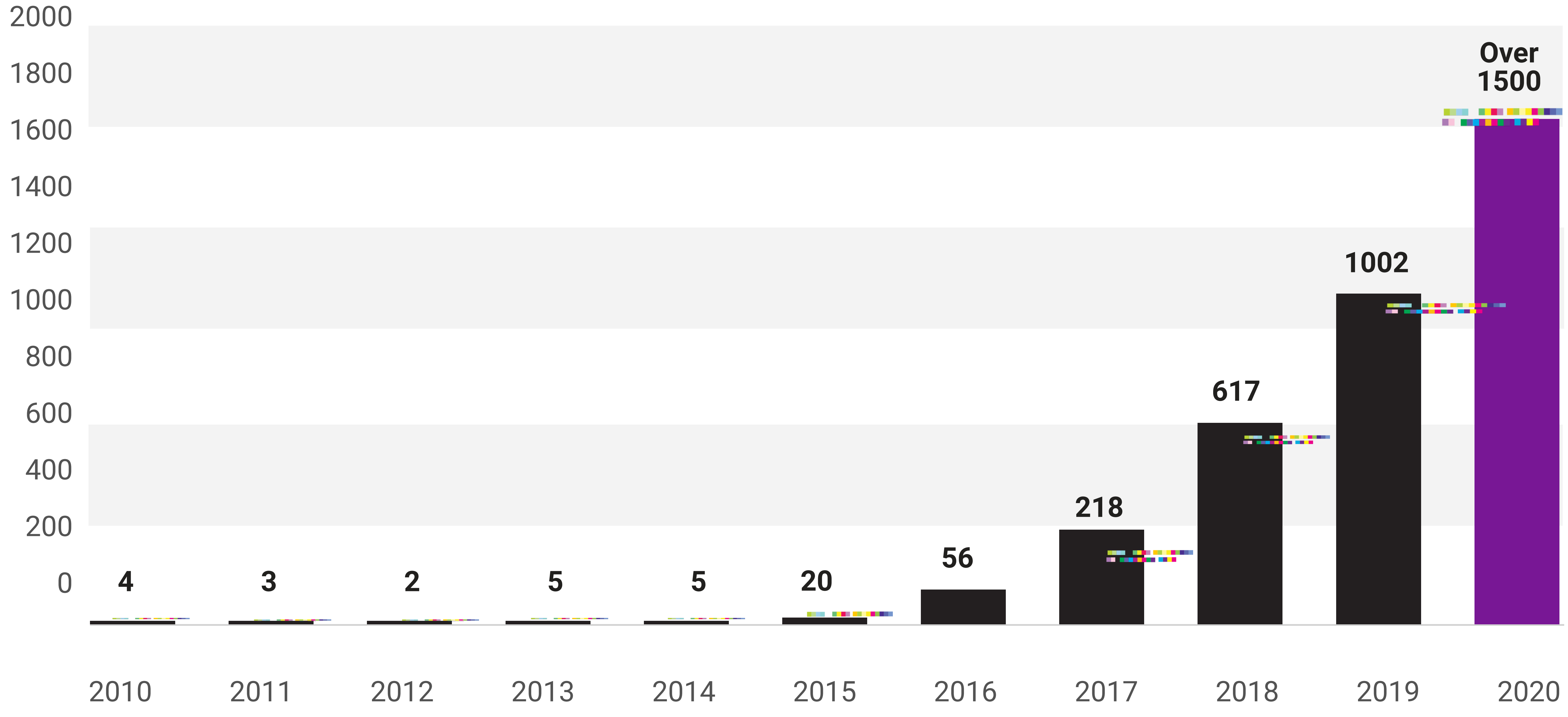
Interest In Security Of AI

A field of practical security for AI seems quite new to many. Others have never heard about it. Security experts have been aware of AI-related data privacy issues and potential attacks on AI systems. However, even for them it is hard to grasp how huge this field is nowadays.

Ideas of adversarial machine learning can be traced as far back as the early 90s. The interest has reborn since 2004, yet it remained a niche area of research. Its fast growth started in 2014 after the publication of "Intriguing properties of neural networks", the first academic paper describing adversarial attacks against deep learning algorithms. Since then the number of research papers has grown significantly (see *Methodology*).

The security of AI emerges, and the number of recent research papers proves this fact.

Number of AI security papers



Takeaways

- For a long time, interest in attacking and defending AI systems has been driven by academia alone. With the AI adoption growth, corporate research labs started to invest in AI security R&D. Lately, governments have caught up with national AI strategies and policies on trustworthy AI, which has spurred even more research.
- More research papers were published in the past two years than in the prior two decades.

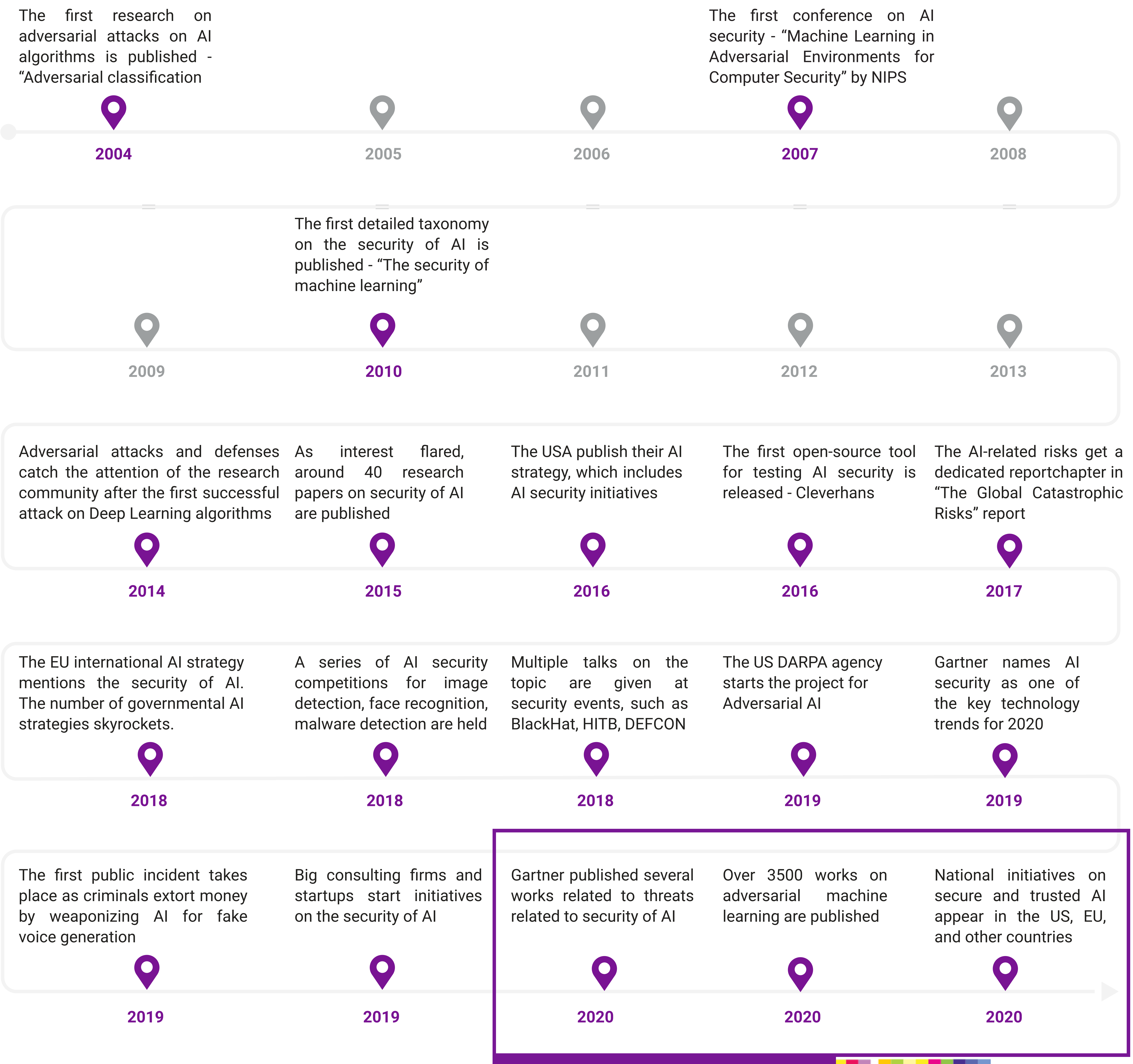
2020 Report Sneak Peek

In 2020, the growth has continued. The increased interest goes in line with commercial AI adoption and governmental initiatives on trusted AI.

Follow us to stay updated and get the new AI security report covering the 2020 year's progress and changes.

Milestones In Security Of AI

We highlight the key events and milestones that have contributed to the development of the field of adversarial machine learning. The rapid growth of interest is shown in the timeline below ranging from the first dedicated research published in 2004 to Gartner's recognition of the AI security as one of the key technological trends for 2020.



The first research on adversarial attacks on AI algorithms is published - "Adversarial classification

2004

2005

2006

The first conference on AI security - "Machine Learning in Adversarial Environments for Computer Security" by NIPS

2007

2008

2009

The first detailed taxonomy on the security of AI is published - "The security of machine learning"

2010

2011

2012

2013

Adversarial attacks and defenses catch the attention of the research community after the first successful attack on Deep Learning algorithms

2014

As interest flared, around 40 research papers on security of AI are published

2015

The USA publish their AI strategy, which includes AI security initiatives

2016

The first open-source tool for testing AI security is released - Cleverhans

2016

The AI-related risks get a dedicated reportchapter in "The Global Catastrophic Risks" report

2017

The EU international AI strategy mentions the security of AI. The number of governmental AI strategies skyrockets.

2018

A series of AI security competitions for image detection, face recognition, malware detection are held

2018

Multiple talks on the topic are given at security events, such as BlackHat, HITB, DEFCON

2018

The US DARPA agency starts the project for Adversarial AI

2019

Gartner names AI security as one of the key technology trends for 2020

2019

The first public incident takes place as criminals extort money by weaponizing AI for fake voice generation

2019

Big consulting firms and startups start initiatives on the security of AI

2019

Gartner published several works related to threats related to security of AI

2020

Over 3500 works on adversarial machine learning are published

2020

National initiatives on secure and trusted AI appear in the US, EU, and other countries

2020

2020 Report Sneak Peek

Takeaways

- The history shows that for various cybersecurity domains it has taken around 10 years to go from the first academic concepts to a complete adoption by organizations. Clearly, the field of secure and trusted AI is on a similar trajectory and is quite close to becoming mainstream, backed by governmental initiatives and industry demand.

Governmental AI Initiatives

There is a concerted effort to develop regulations and ethical guidelines for AI systems. On June 29 2019, G20 leaders signed a statement on basic ethical principles for AI. The OECD.AI Policy Observatory provides data on AI's areas of impact. With the first timeline we show the declarations on AI made by individual countries, and the second one illustrates initiatives focusing on the secure and trusted AI that preceded the international consensus.

October 2016
USA
The National Artificial Intelligence Research And Development Strategic Plan: 2016

March 2017
Canada
Pan-Canadian Artificial Intelligence Strategy

March 2017
Japan
Artificial Intelligence Technology Strategy

April 2017
China
Artificial Intelligence: Implications for China

May 2017
Singapore
AI Singapore

October 2017
UAE
UAE Strategy for Artificial Intelligence

November 2017
UK
Interim Cyber Security Science & Technology Strategy

December 2017
Finland
Finland's Age of Artificial Intelligence

January 2018
Taiwan
AI Taiwan

January 2018
Saudi Arabia
Vision 2030

January 2018
Qatar
National Artificial Intelligence Strategy For Qatar

January 2018
Germany
Artificial Intelligence Strategy

March 2018
France
For a Meaningful Artificial Intelligence

March 2018
Italy
Artificial Intelligence at the Service of Citizens

April 2018
EU
Communication Artificial Intelligence for Europe

May 2018
Sweden
National Approach for Artificial Intelligence

June 2018
Australia
Artificial intelligence and privacy

June 2018
India
National Strategy for Artificial Intelligence

June 2018
Mexico
Towards an AI Strategy in Mexico

January 2019
Poland
Map of the Polish AI

March 2019
Denmark
National Strategy for Artificial Intelligence

July 2019
The Czech Republic
National Artificial Intelligence Strategy

October 2019
Russia
National Strategy for the Development of Artificial Intelligence by 2030

December 2019
The Republic of Korea
National Strategy for Artificial Intelligence

February 2019
USA
Secure, Assured, Intelligent Learning Systems and Trojans in AI

April 2019
EU
The Ethics Guidelines for Trustworthy Artificial Intelligence

May 2019
Australia
Artificial Intelligence: Australia's Ethics Framework

June 2019
China
Governance Principles for a New Generation of Artificial Intelligence

June 2019
EU
Policy and Investment Recommendations for Trustworthy Artificial Intelligence

October 2019
USA
AI Principles: Recommendations on the Ethical Use of AI

2020 Report Sneak Peek

In 2020, new AI strategies and security-focused initiatives appeared in Europe and other parts of the world.

Follow us to stay updated and get the new AI security report covering the 2020 year's progress and changes.

Takeaways

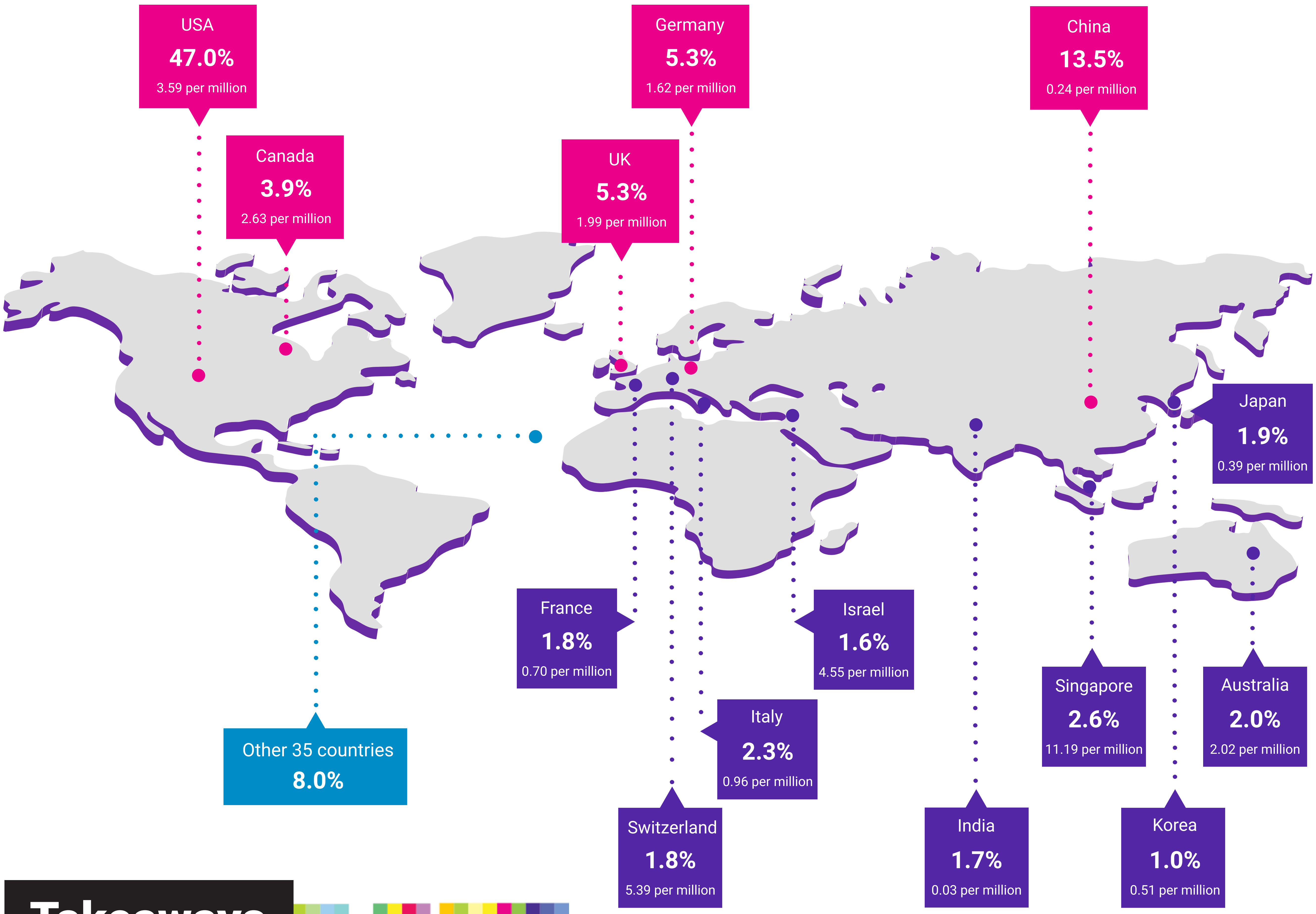
- The United States was first to confirm a critical role of AI for governments with their national AI strategy. It took significant time for other countries to catch up with their AI initiatives.
- The first initiatives on trustworthy AI appeared in 2019. It took around 2 years to shift focus from the basic AI policies to the trusted AI, so the trend will certainly continue.

AI Security Research By Country

In total, 49 countries have contributed to the field of adversarial machine learning. Most researchers are affiliated with academic and industrial labs. 77% of papers are written by authors from the same country while collaborations involving two or more countries take up the shares of 20% and 3% respectively.

The volume of security research is correlated to the maturity of tech industries. Predictably, the USA and China, two AI superpowers, have published more papers than the rest of the world combined. The UK, Germany, and Canada, some of the world's largest economies with a proven track record for innovation, are far behind but the blooming interest in AI adoption is motivating them to catch up.

The map reveals disparities in AI security R&D investment among countries which contributed at least 1% to a total number of research publications (see *Methodology*).


USA 47.0% 3.59 per million
Germany 5.3% 1.62 per million
China 13.5% 0.24 per million
Canada 3.9% 2.63 per million
UK 5.3% 1.99 per million
Japan 1.9% 0.39 per million
France 1.8% 0.70 per million
Israel 1.6% 4.55 per million
Singapore 2.6% 11.19 per million
Australia 2.0% 2.02 per million
Italy 2.3% 0.96 per million
Other 35 countries 8.0%
Switzerland 1.8% 5.39 per million
India 1.7% 0.03 per million
Korea 1.0% 0.51 per million

Takeaways

- Based on the total number of publications, the United States and China predictably lead, while the United Kingdom and Germany share third place. Although, if we consider research per capita (in millions), the top 3 countries are Singapore, Switzerland, and Israel.
- Due to the siloed research environment and language differences with China, we expect that their real number of research publications and state-sponsored AI security research and development is significantly higher than our research methodology is able to show.

Coverage Of AI Security Areas

Ensuring security of AI systems is complicated. AI-powered software is significantly harder to develop and maintain than traditional software. Still, even traditional application security is not yet a solved problem. Being non-deterministic, AI systems tend to exhibit different behaviors in the same conditions, which makes it even more difficult to protect.

To develop trusted AI, security practitioners should pay equal attention to all stages of the AI development lifecycle including threat awareness, risk assessment, and security operations. Here we show how research interest is distributed across AI security areas of concern.

Surveys

2.6% of papers review AI security progress

To understand the complex field of adversarial machine learning and raise security awareness across AI stakeholders, researchers review AI security threats, investigate existing attacks and defenses, and explore paths to trustworthy AI through making AI systems resilient.

Attacks

49.0% of papers analyze attacks against AI algorithms

To conduct a risk assessment of AI systems, researchers invent new ways of hacking AI models. Others exploit AI infrastructure, side channels, and software bugs. These methods can be compared to general static or dynamic application security testing as well as traditional penetration testing.

Defenses

47.2% of papers suggest defenses for AI algorithms

To protect against AI model attacks, researchers suggest transformations of data inputs and outputs, algorithm modifications, and secure model retraining. Less popular options include input verification or anomaly detection. These methods correspond with vulnerability remediation and threat detection.

Tools

1.2% of papers introduce tools for AI security assurance

To implement security operations and ensure the resilience of AI systems, researchers develop tools for automating vulnerability testing and verification. Such tools look similar to common vulnerability scanners, yet they are far from ready to be productionized and used outside research labs.

Takeaways

- Close numbers for attacks and defenses for AI reflect the common cat-and-mouse game in cybersecurity. So, adversarial machine learning follows the same cybersecurity laws.
- Even though the number of attack and defense papers is quite close, there are dozens of effective attacks and almost no efficient defenses for AI systems.

Categories Of Attacks On AI

The attack surface of AI systems is huge. This allows cybercriminals to attack multiple layers of the AI ecosystem. The base layer is generic or specialized AI hardware. On top of this is either on-premises or cloud infrastructure. This then hosts a software technology stack running the application's code which also integrates machine learning frameworks for AI algorithms. These algorithms are then continuously fed inputs through data pipelines.

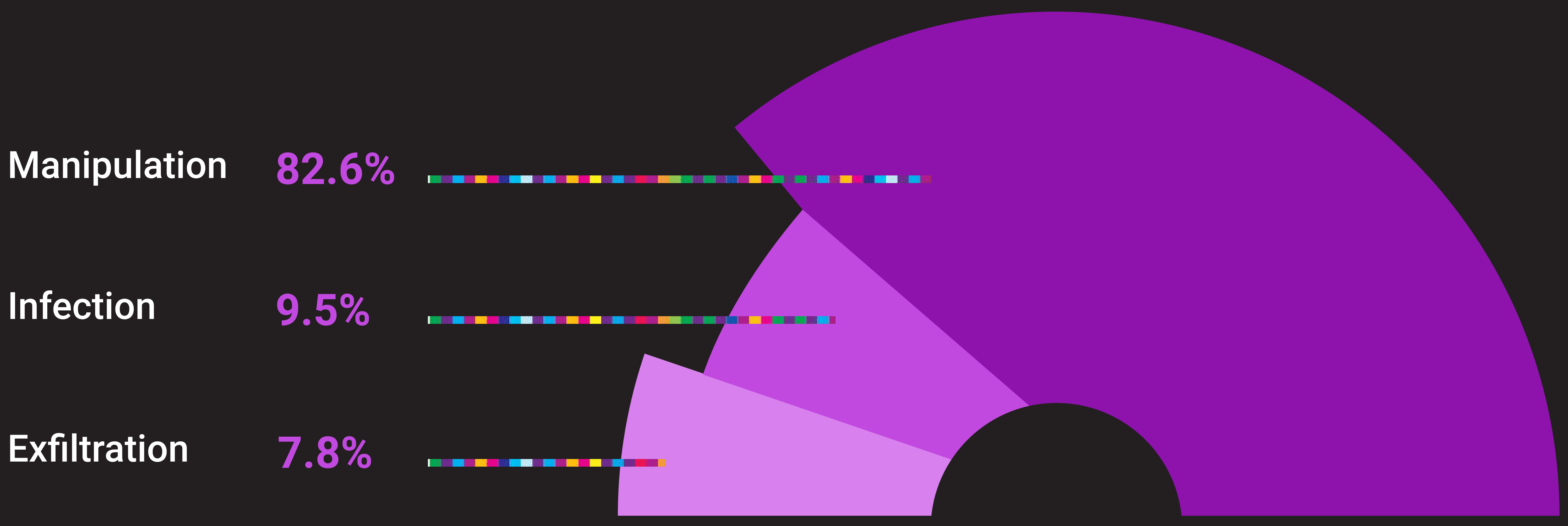
Clearly, there are many variables weakening the security of companies' AI crown jewels. This report focuses on the last frontier of AI security represented by core AI functionality – machine learning algorithms, the main focus of the adversarial machine learning field.

In this section, we show the most mature categories of adversarial attacks.

- **Manipulation attacks** allow adversaries to bypass expected AI behavior or even make AI systems perform unexpected jobs. With maliciously crafted inputs attackers can conduct evasion attacks or adversarial reprogramming of AI systems in real-time.

- **Infection attacks** sabotage the quality of AI decisions and enable stealth control of AI systems. Attackers contaminate data used for training, exploit hidden triggers in AI behaviors, or distribute malicious AI models via poisoning, trojan, or backdoor attacks.

- **Exfiltration attacks** aim to steal data from AI systems. Data samples used for AI training, private AI inputs, internals of AI algorithms can be exfiltrated with attacks such as model inversion, membership and attribute inference, or model extraction.



Takeaways

- Data privacy and intellectual property theft are critical problems in cybersecurity. However, the interest in data exfiltration and extraction of AI system's internals appears to be quite low. Such disbalance can mislead into thinking that exfiltration risks are not important.

Top 10 Attacks On AI Systems

1	**Evasion attack** bypasses normal decisions by AI systems in favor of attacker-controlled behavior by crafting malicious data inputs called adversarial examples	81.0%	Manipulation
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	6.8%	Infection
3	**Membership inference attack** discloses whether specific data sample was a part of a dataset used for AI training	3.5%	Exfiltration
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	2.3%	Infection
5	**Model extraction attack** exposes algorithm's internal details by making malicious queries to AI systems	1.9%	Exfiltration
6	**Attribute inference attack** reveals secret data details by exploiting public information received from AI systems	1.3%	Exfiltration
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	1.2%	Infection
8	**Model inversion attack** reveals secret data inputs based on public outputs by maliciously querying AI systems	1.2%	Exfiltration
9	**Anti-watermarking attack** bypasses protection controls used by AI systems for copyright or authenticity checks	0.6%	Exfiltration
10	**Reprogramming attack** allows threat actors to repurpose AI models and make them execute unexpected tasks	0.2%	Manipulation

Takeaway

- Most attacks against AI systems resemble traditional application security attacks. It's traceable how attack vectors are methodically taken from general cybersecurity and applied to adversarial machine learning. So, some traditional attacks are yet to be adopted.

2020 Report Sneak Peek

In 2020, attack surface has expanded to include attacks against AI system's software and hardware environments.

Follow us to stay updated and get the new AI security report covering the 2020 year's progress and changes.

Lifecycle For Security Of AI

To deal with new threats covered in this paper, we suggest the following Secure AI Lifecycle to jumpstart an AI security program. The four areas represent the AI system's stages from inception to maturity and looped back for continuous improvement. The steps are ordered from basic to complex with later steps relying on preceding outcomes. This lifecycle corresponds with NIST Cybersecurity Framework and Gartner's Adaptive Security Architecture, popular reference frameworks for cybersecurity lifecycle management.

Identify 1

Goal

Understand current AI security posture with asset management, threat modeling, and risk assessment activities

Steps

1. **Asset management**: Identify and document all used AI models, datasets, cloud platforms and vendors
2. **Threat modeling**: Understand risks of compromising AI models, datasets, their environments and supply chains
3. **Risk assessment**: Perform a security audit and prioritize vulnerabilities in AI models, datasets, and their environments

Protect 2

Goal

Implement protective controls such as security awareness, system hardening, and practices for secure AI development

Steps

1. **Security awareness**: Educate stakeholders from leadership, product security, and AI development about security risks
2. **Model hardening**: Apply security defenses against attacks on AI models, ensure safe inputs, prevent data exfiltration
3. **Secure development**: Establish a regular process for AI application security covering a pipeline from development to production

Detect 3

Goal

Defend against active adversaries with security monitoring and threat detection systems validated by regular penetration testing

Steps

1. **Security monitoring**: Collect and analyze events and anomalies from production AI systems such as access, errors, and metrics issues
2. **Threat detection**: Detect and block adversarial attacks targeting AI system confidentiality, integrity, and availability
3. **Penetration testing**: Conduct red team excercises to assess adversarial robustness and check detection and response controls

Respond 4

Goal

Prepare for AI security incidents by introducing investigation, containment practices, mitigation tools, techniques, and procedures

Steps

1. **AI forensics**: Build an expertise for AI security incident classification, impact analysis, and technical investigation
2. **Incident response**: Develop playbooks for incident containment and communication with stakeholders
3. **Mitigation**: Improve technical controls and organizational policies to reduce chances of repeated AI security incidents

Closing Words

AI has demonstrated clear benefits for the modern world, so the adoption of AI will continue to grow. As more devices, platforms, and applications are powered by AI algorithms, it's crucial to protect them from cyber threats, privacy issues, and safety incidents.

Due to fundamental differences, traditional cybersecurity products don't work for securing AI. Unfortunately, reliable AI-focused security solutions have not been developed yet. This makes the AI industry terrifyingly unprepared for real-world attacks against AI systems.

This report only touches the tip of the iceberg. Building a trustworthy AI takes a collaborative effort from various communities focused on traditional cybersecurity, fairness, accountability, transparency, ethics, privacy, security, and safety of AI.

Yet, this requires support and investment from business stakeholders. Many of them are not aware of these challenges. This report aims to raise awareness of the trusted AI problem.

Only together, can researchers and businesses bridge the gap between theory and practice of operationalizing trustworthy AI, to make the AI revolution happen.

Methodology

Here we explain how we obtained, analyzed, and chose to present information in this report.

AI industry voice. Every quote in the section was provided by an industry expert after they read this report or when we reached out to them for their commentary on the topic.

Progress toward secure AI. The information to summarize progress in academia, industry, and public initiatives was obtained from open sources. We chose what to highlight and formulated predictions based on the expert opinions of our team.

AI security incidents. The information to showcase real-world incidents was obtained from open sources. We chose what to highlight based on the expert opinions of our team.

AI under attacks chapters. In these chapters, we present the results of our analysis of the corpus of security research related to artificial intelligence and machine learning.

We focus on practical attacks and defenses for AI systems, the field called adversarial machine learning. We keep out of scope papers which do not imply malicious intentions of threat actors. Works used adversarial ML techniques in non-malicious contexts such as adversarial examples for explainability or adversarial training for general robustness were excluded. We excluded papers about the general robustness of safety-critical AI systems or control problems of AI superintelligence. We also excluded papers on fairness, accountability, and transparency unless research covers attacks or defenses for algorithms of trusted AI.

Our main goal was to analyze trends and research interest in the security for AI rather than evaluate peer-reviewed academic research from publications and conferences. Thus, we have selected arXiv.org as the biggest open-access collection of non-peer-reviewed research. We have only used other resources for cross-validation and enrichment of our original dataset. Our dataset includes 1932 relevant papers published during the 2010 - 2019 period.

AI areas under attacks. We grouped attacked AI applications based on the AI areas they pertain to. Vision embraces computer vision applications, language involves natural language processing and speech recognition applications, analytics includes applied data analysis across multiple industries, and autonomy covers reinforcement learning applications.

AI datasets under attacks. We determined the types of datasets used by AI models, which were attacked by researchers.

AI applications under attacks. We determined AI applications based on the practical tasks of attacked models and datasets.

AI industries under attacks. We determined industries based on the purpose of attacked models and datasets. For instance, attacks against AI models for semantic segmentation with datasets from car dashboards' cameras were counted towards the automotive industry.

Methodology

Interest in security of AI. We counted all adversarial machine learning research papers by years from 2010 till 2019. The date that a paper is initially submitted to arXiv.org is considered to be the "publication date". The rule stands even if research had been published in other journals before or was resubmitted later. Data for 2020 year is provided as a sneak peek of the follow-up report and may change later.

Milestones in security of AI. We defined milestones as events connected to the security of AI that were widely publicized and have significantly affected the industry. We decide to present milestones that are most noteworthy based on our expert opinion. The timeline range includes the decade from 2010 till 2019. Data for 2020 year is provided as a sneak peek of the follow-up report and may change later.

Governmental AI initiatives. We showed the timeline based on the corpus of public documents that detail a national strategy to develop and regulate AI and Security of AI. Only those documents produced by a single governing body were included. We have omitted papers produced by commercial and academic organizations. We admit that some works and industry-specific initiatives might have been missed.

The collected documents were sorted by year, country, and purpose to provide a better understanding of strategic plans and ongoing efforts related to AI, or its security and safety. Only those initiatives that deal with AI system security are presented within this section. For each year between 2010 and 2019 we chose the most significant document to represent on the timeline. Data for 2020 year is provided as a sneak peek of the follow-up report and may change later.

AI security research by country. We attributed research papers to a particular country based on the author's self-reported affiliation to an institution and its headquarters' location. If a regional office was clearly identified, it's location was used. In case the author was affiliated with several international institutions, the research counted towards all of them. Per capita numbers are counted in millions for countries with at least 1% contribution.

Coverage of AI security areas. We classified all research papers by areas such as "survey", "attack", "defense", "toolbox".

Categories of attacks on AI. We classified all attacks by categories. This classification is based on high-level group each adversarial attack method belongs to.

Top 10 attacks on AI systems. We identified and counted all attacks cases to comprise the list in the descending order. Data for 2020 year is provided as a sneak peek of the follow-up report and may change later.

Lifecycle for security of AI. We adopted NIST Cybersecurity Framework and Gartner's Adaptive Security Architecture for secure AI lifecycle. The four categories represent AI system lifecycle from inception to maturity. The steps are ordered from basic to complex with later steps relying on preceding outcomes.

ADVERSA is an award-winning Israeli AI research company on a mission to increase trust in AI systems by protecting them from cyber threats, privacy issues, and safety incidents.

We unite world-class cybersecurity experts, mathematicians, AI researchers, and neuroscientists who work together to shape the AI industry, develop innovative methodologies, frameworks, and solutions for Secure and Trusted AI.

Authors, Alex Polyakov and Eugene Neelou, lead their research teams in conducting a deep analysis of industry, academic, and governmental signals to shed light on one of the most critical AI problems of the next decade.

We invite enthusiasts, researchers and industry partners to join us on the **Road to Secure and Trusted AI**.

Contact us: info@adversa.ai
Subscribe for updates: adversa.ai/stay-updated
Get the latest report version: adversa.ai/report-secure-and-trusted-ai

First published on 21-April-2021