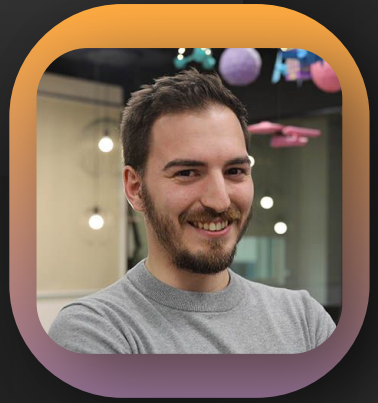
/RootedCON

Inteligencia Artificial Ofensiva

¿Cómo podemos estar preparados?

Miguel Hernández
José Ignacio Escribano



Miguel Hernández

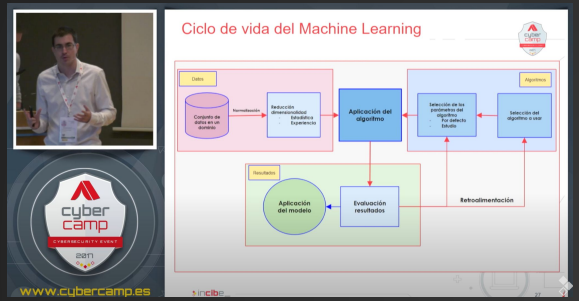
Security Content Engineer at sysdig
@MiguelHzBz @mastodon.social

José Ignacio Escribano

Cybersecurity and Machine Learning Researcher
/in/josé-ignacio-escribano-pablos

Anteriormente en AI + Seguridad...

2017



https://www.youtube.com/watch?v=pTzwTTVflZs

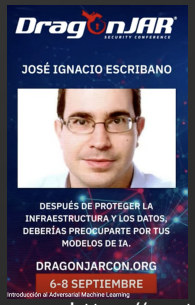
2019

https://www.youtube.com/watch?v=ist4Za3C2DY&ab_channel=CCN

2021

https://www.youtube.com/watch?v=jsDLt5gBnmY&ab_channel=CCN

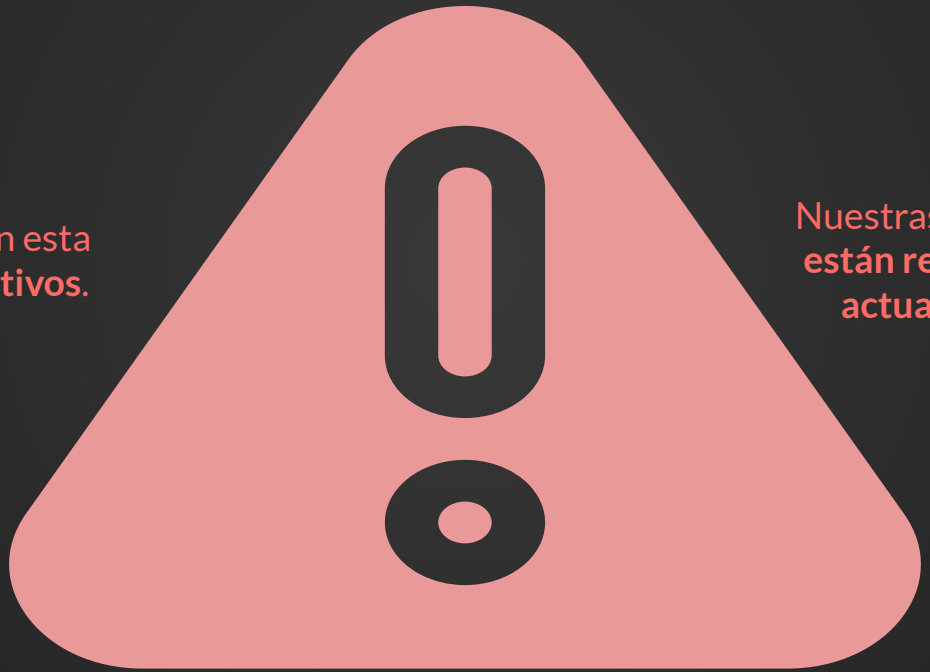
2022



https://youtu.be/SzW6E6d8KOs?t=27358

Disclaimer

Todo lo que se muestra en esta charla es con fines **educativos**.



Nuestras opiniones personales **no están relacionadas con nuestros actuales centros de trabajo**.

Agenda

01
¿Qué es IA ofensiva?
Motivaciones, límites…

02
Abuso
Adversarial Machine Learning

03
Uso
Mejorando lo existente

04
¿Cómo estar preparados?
Recomendaciones
y recursos útiles

01

¿Qué es la IA Ofensiva?

IA ofensiva

La IA ofensiva es el uso de inteligencia artificial con un propósito malicioso que incluye:

- Ataque a sistemas de inteligencia artificial (o Adversarial machine learning)
- Ataques “clásicos” mejorados con inteligencia artificial (persona building, malware defense evasión, y demás que veremos en esta presentación).

“The use or abuse of AI to accomplish a malicious task”

https://arxiv.org/pdf/2106.15764.pdf

Motivación del adversario

Automatización

Escalado automático de tareas complejas (spear phishing attacks).

Velocidad

Mayor velocidad en alcanzar las metas del adversario.

Éxito

Aumento en la probabilidad de éxito.

Nuevas amenazas

Propaganda

Posibilidad de realizar ataques personalizados al alcance de cualquiera. Por ejemplo, IA generativa, etc.

Malware inteligente

Si el malware es capaz de hacer movimientos laterales/discovery, reducirá la comunicación con C2 y será más difícil detectarlo.

Robo de propiedad intelectual

La IA permite el robo de la propiedad intelectual (Adversarial Machine Learning).

Taxonomía

Abuso

Explotar las vulnerabilidades de los modelos de Inteligencia Artificial.

Uso

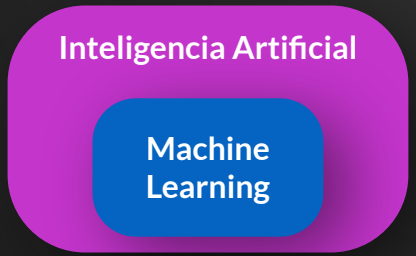
Mejorar las técnicas ofensivas actuales con el uso de la Inteligencia Artificial.

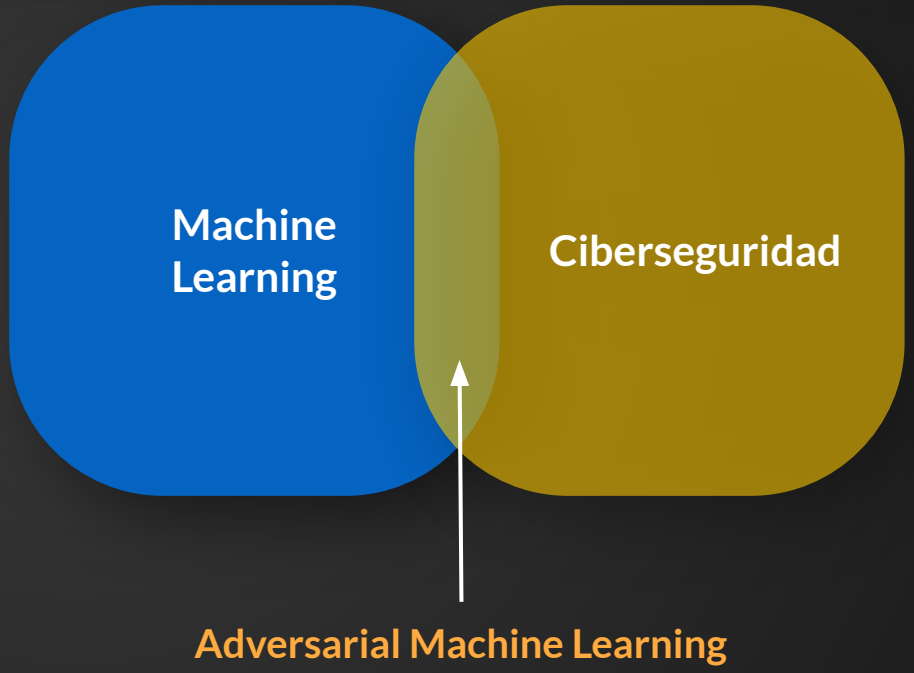
02 Abuso

Adversarial Machine Learning

Adversarial Machine Learning

Rama del machine learning que estudia los **ataques** que puede sufrir un **modelo en la presencia de un adversario malicioso** y cómo **protegerse** de ellos.





Taxonomía

Extracción

Robo de modelos

Inversión

Filtrado de datos

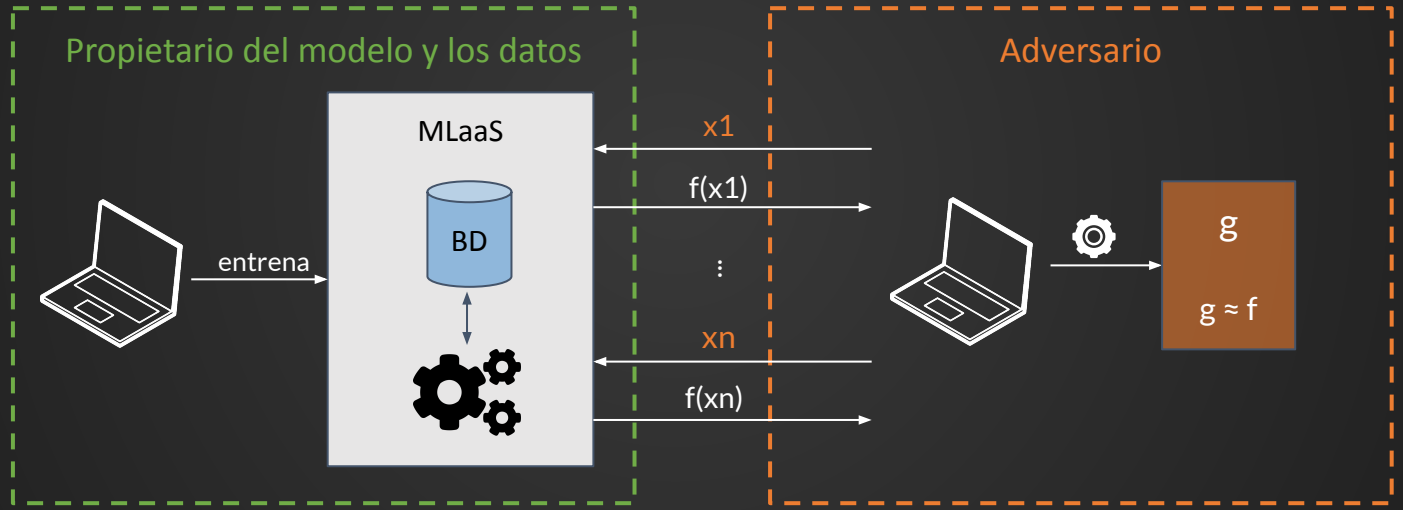
Envenenamiento

Creación de puertas traseras

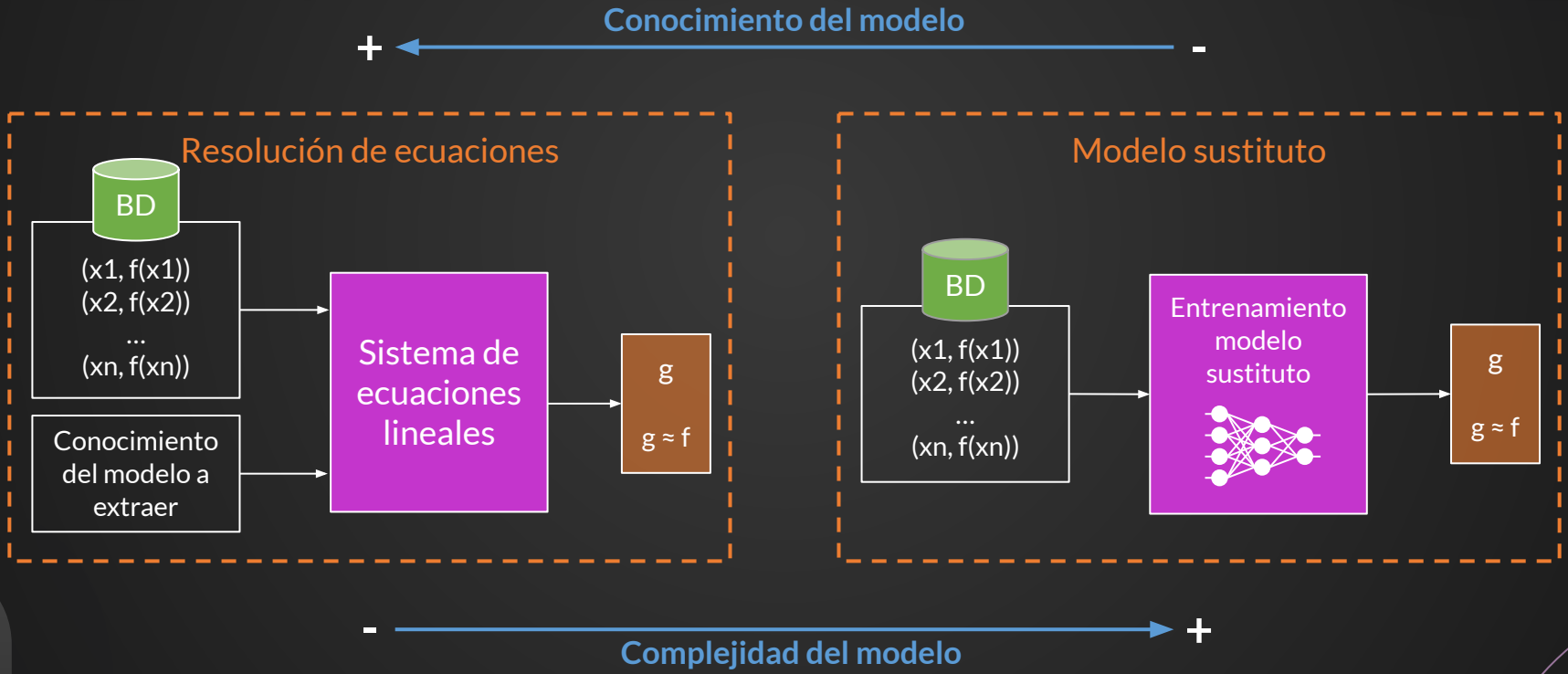
Evasión

Engaño a los modelos

Extracción (o robo de modelo)



Extracción (o robo de modelo)



Extracción (o robo de modelo)

Limitaciones

El ataque no es sencillo en entornos reales.

Peticiones limitadas por el modelo.

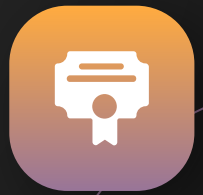
En ocasiones, es tan complicado como entrenar desde cero.

Defensas

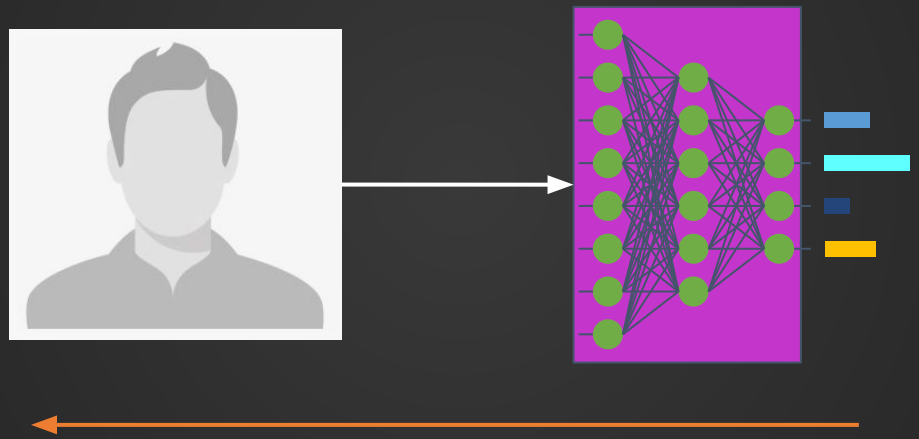
Redondeo de las salidas.

Privacidad diferencial.

Medidas específicas.



Inversión



Inversión

Membership Inference Attack (MIA)

Determinar si una **muestra** fue empleada como **parte del entrenamiento.**

Property Inference Attack (PIA)

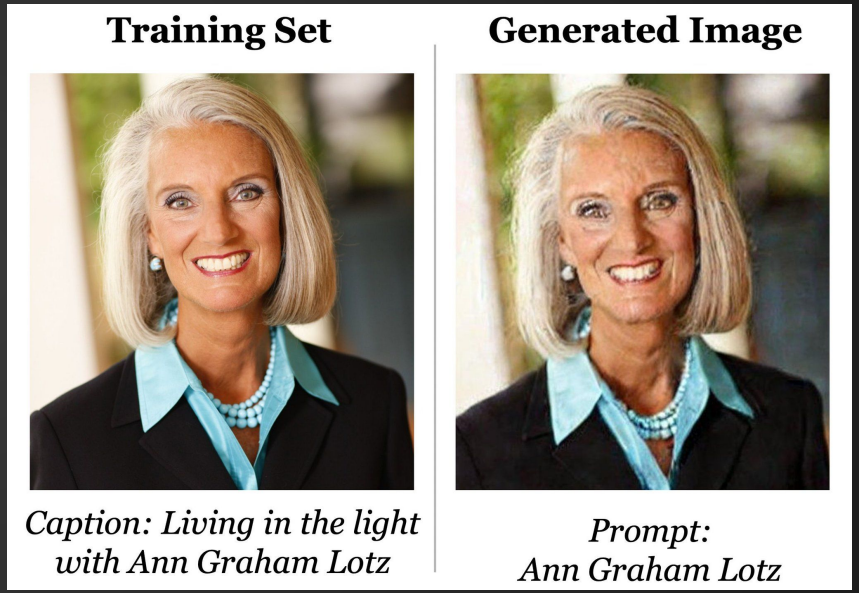
Extracción de **propiedades estadísticas** que no fueron **codificadas** durante la **fase de entrenamiento.**

Reconstrucción

Recreación de una o más **muestras del conjunto de entrenamiento** y/o sus **etiquetas**.

Inversión

https://dl.acm.org/doi/10.1145/2810103.2813677



https://arxiv.org/abs/2301.13188

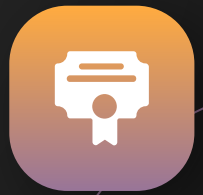
Inversión

Defensas

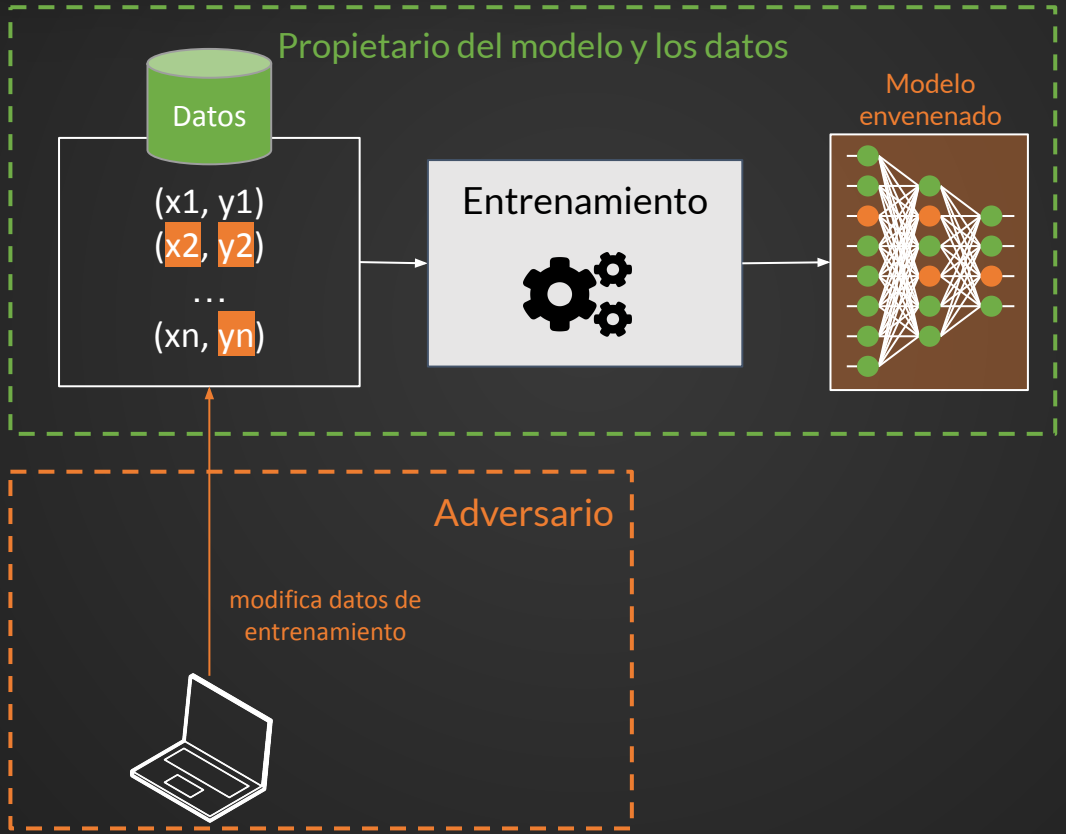
Criptografía avanzada como privacidad diferencial, criptografía homomórfica y computación multiparte segura.

Técnicas de regularización como Dropout.

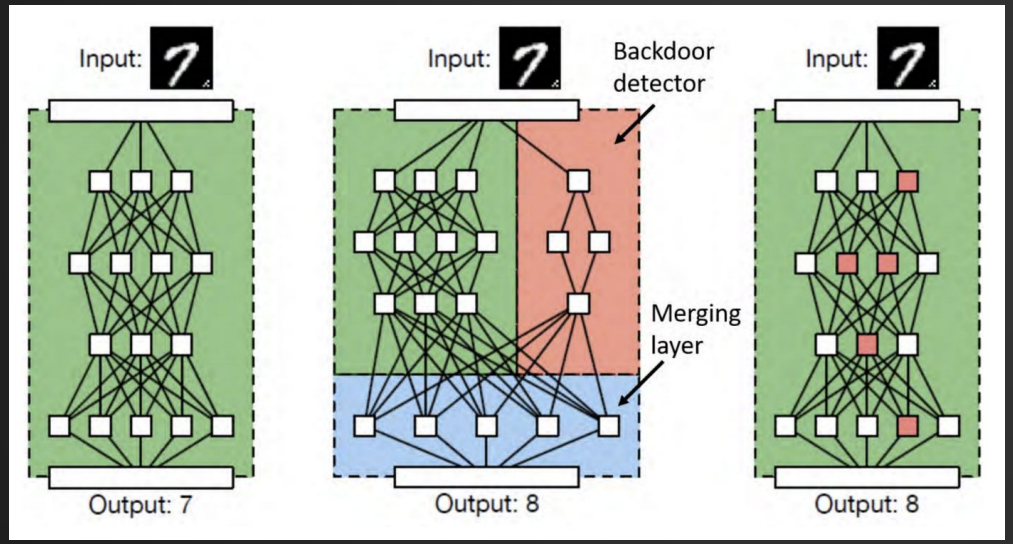
Compresión de modelos.

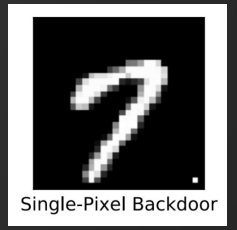


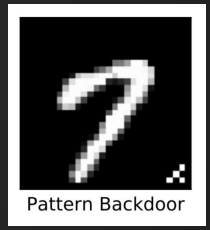
Envenenamiento



Envenenamiento







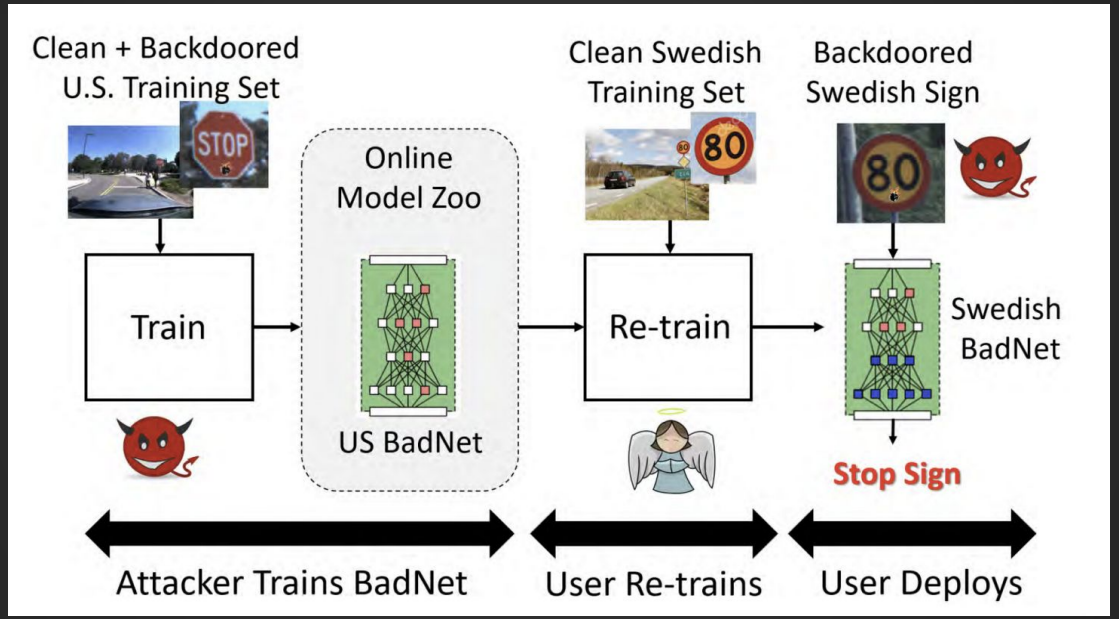
https://dl.acm.org/doi/10.1145/2810103.2813677

Envenenamiento

Imágenes reales

https://dl.acm.org/doi/10.1145/2810103.2813677

Envenenamiento



https://dl.acm.org/doi/10.1145/2810103.2813677

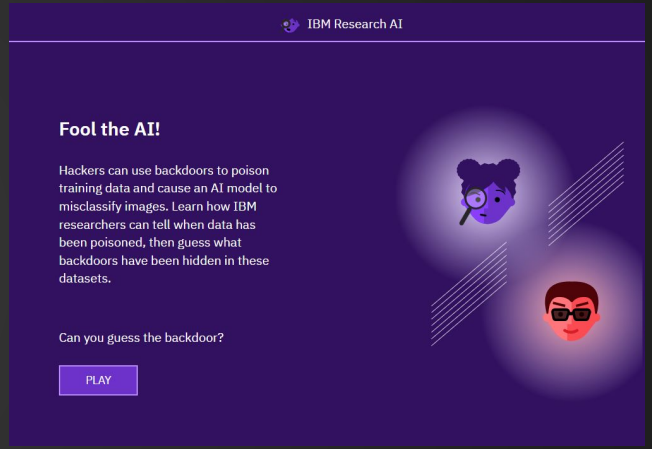
Envenenamiento

Defensas

Protección del dato, que incluye evitar la modificación, la denegación y la falsificación de los datos y, la detección de los datos envenenados, junto con el uso del saneamiento de datos.

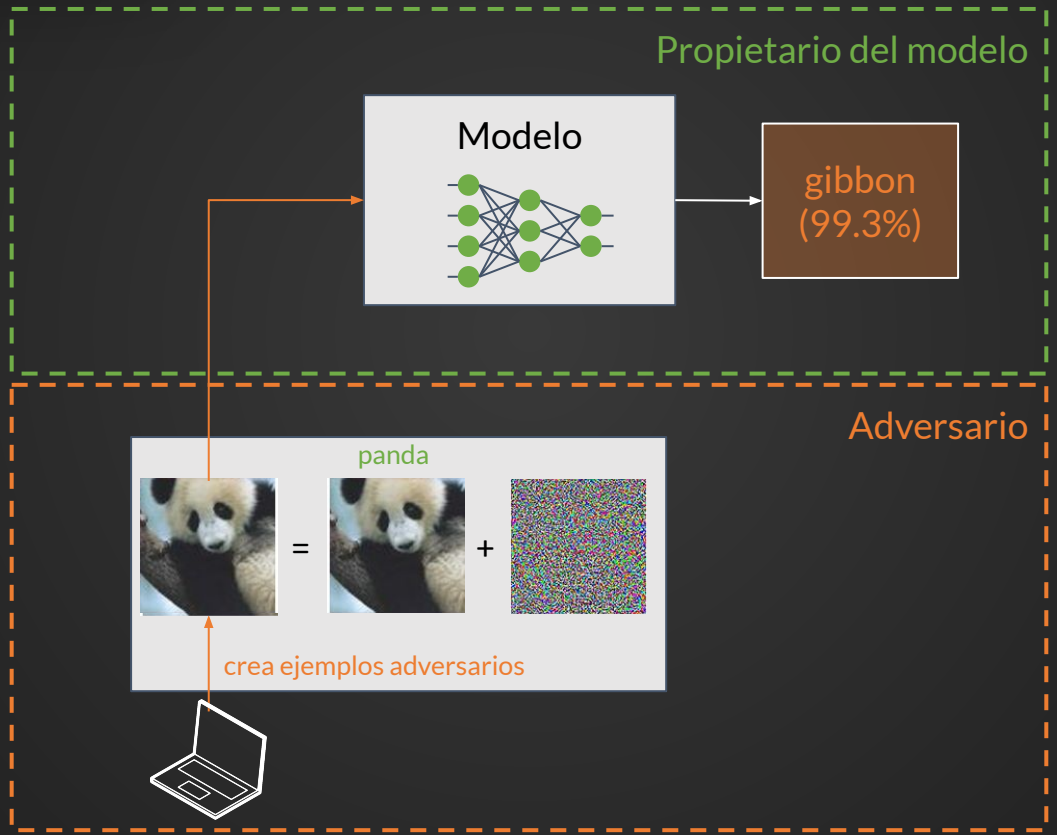
Protección de los algoritmos, que intenta emplear métodos robustos de entrenamiento.

Defensas específicas.

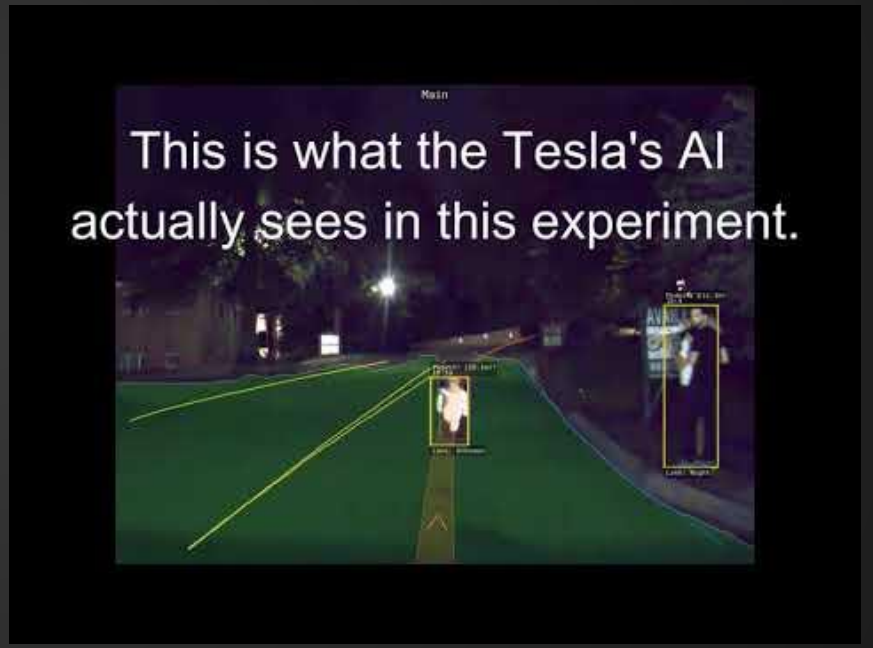


https://fooltheai.mybluemix.net

Evasión

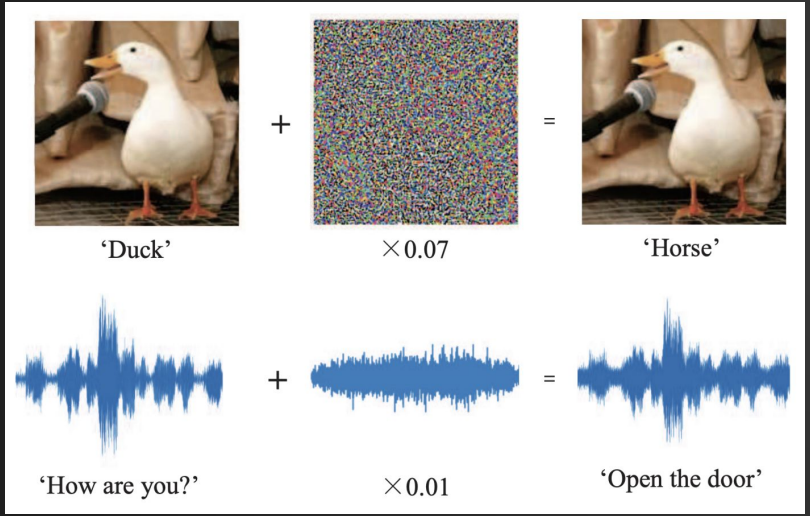

Propietario del modelo
Modelo
gibbon
(99.3%)
Adversario
panda
=
+
crea ejemplos adversarios

Evasión



https://www.nassiben.com/phantoms

Evasión

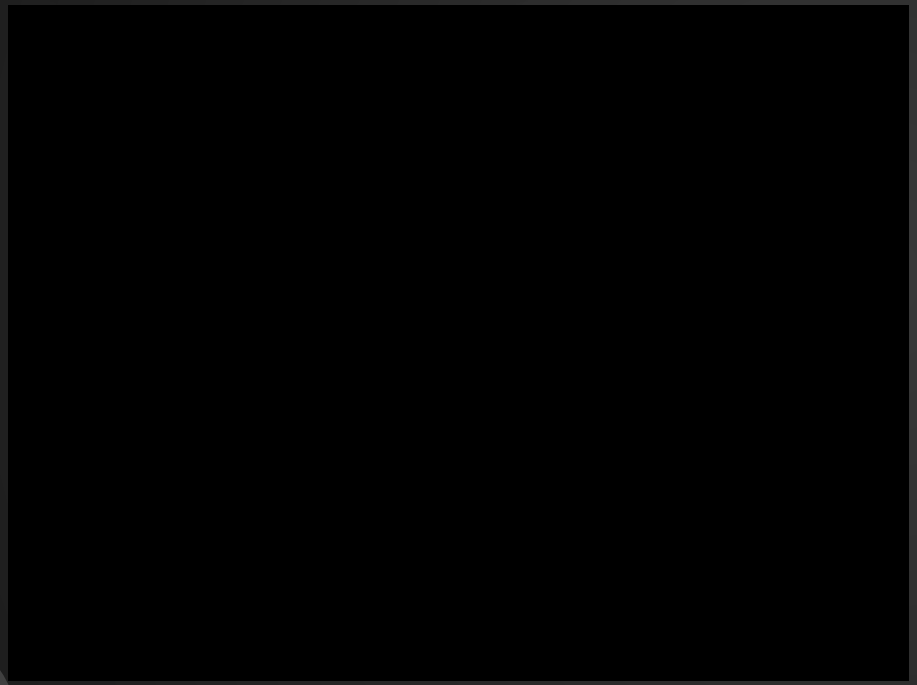


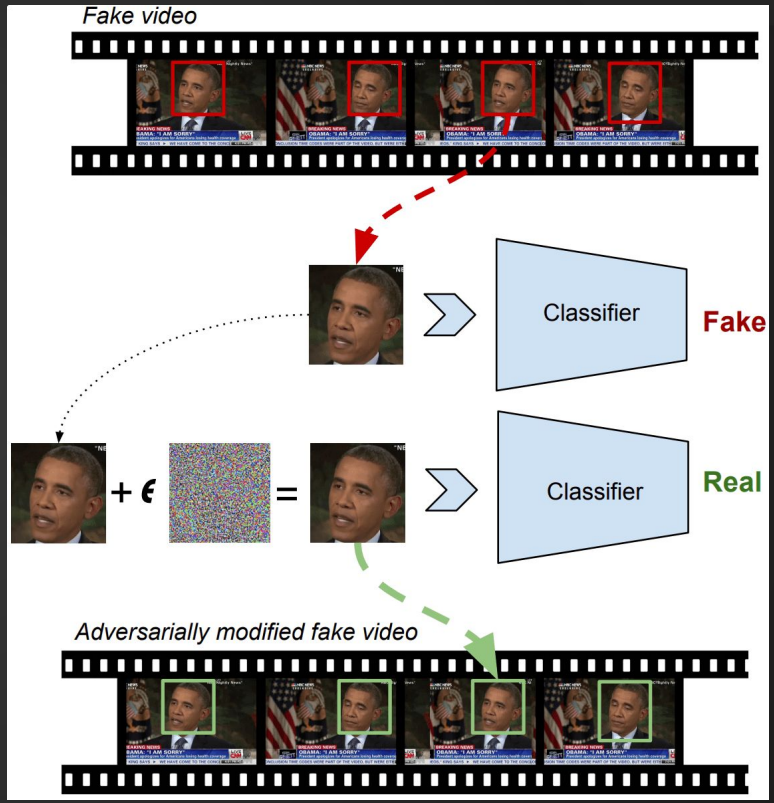
Classifier: Word-LSTM **Original Text Prediction**: Sci/Tech (Confidence = 47.80%) **Adversarial Text Prediction**: Business (Confidence = 52.52%)
Original Text: *Tyrannosaurus rex achieved its massive size due to an enormous growth spurt **during** its adolescent years.* **Adversarial Text**: *Tyrannosaurus rex achieved its massive size due to an enormous growth spurt **durnig** its adolescent years.*
Classifier: Text-CNN **Original Text Prediction**: Company (Confidence = 98.16%) **Adversarial Text Prediction**: Artist (Confidence = 20.27%)
Original Text: *Yates is a gardening company **in** New Zealand and Australia.* **Adversarial Text**: *Yates is a gardening company **i** New Zealand and Australia.*

https://arxiv.org/abs/1803.09156

https://arxiv.org/abs/2002.00760

Evasión





https://adversarialdeepfakes.github.io

Evasión

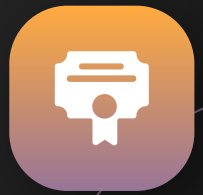
Defensas

Entrenamiento adversario, que consiste en emplear ejemplos adversarios durante el entrenamiento para que el modelo aprenda características de los ejemplos adversarios, haciendo más robusto el modelo ante este tipo de ataque.

Transformaciones sobre las entradas.

Enmascarado/regularización del gradiente. No muy efectiva.

Defensas débiles.



Herramientas open source

Nombre	Tipo	Algoritmos	Tipos de ataques	Ataque/Defensa	Frameworks soportados
Cleverhans	Imagen	Deep Learning	Evasión	Ataque	Tensorflow, Keras, JAX
Foolbox	Imagen	Deep Learning	Evasión	Ataque	Tensorflow, PyTorch, JAX
ART	Cualquiera	Deep Learning, SVM, LR, etc.	Todos	Ambos	Tensorflow, Keras, Pytorch, Scikit Learn
TextAttack	Texto	Deep Learning	Evasión	Ataque	Keras, HuggingFace
Advertorch	Imagen	Deep Learning	Evasión	Ambos	----
AdvBox	Imagen	Deep Learning	Evasión	Ambos	PyTorch, Tensorflow, MxNet
DeepRobust	Imagen, grafos	Deep Learning	Evasión	Ambos	PyTorch
Counterfit	Cualquiera	Cualquiera	Evasión	Ataque	----
Adversarial Audio Examples	Audio	DeepSpeech	Evasión	Ataque	----

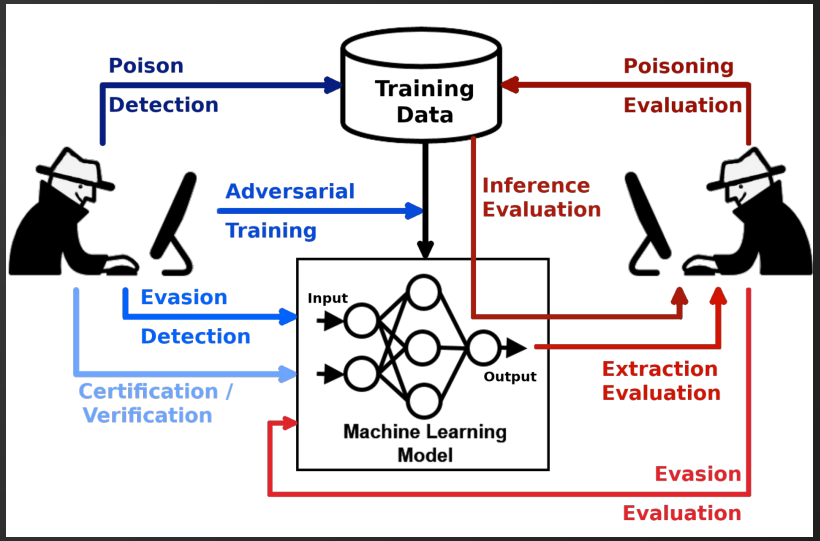
Principales herramientas


```
-------------------------------------------------
                  Microsoft
                                  __          _____ __
  _________  __  ______  / /____  _____/ __(_) /_
 / ___/ __ \/ / / / __ \/ __/ _ \/ ___/ /_/ / __/
/ /__/ /_/ / /_/ / / / / /_/  __/ /  / __/ / /
\___/\____/\__,_/_/ /_/\__/\___/_/  /_/ /_/\__/

                                        #ATML

-------------------------------------------------
```

Adversarial Robustness Toolbox (ART)



https://github.com/Trusted-AI/adversarial-robustness-toolbox

03
Uso ofensivo

Mejorando lo existente

Algunas aplicaciones

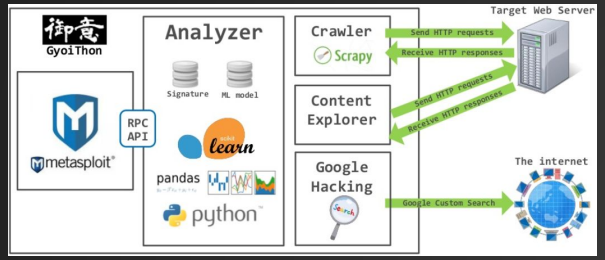
IA Potenciando Pentesting

IA Potenciando Malware

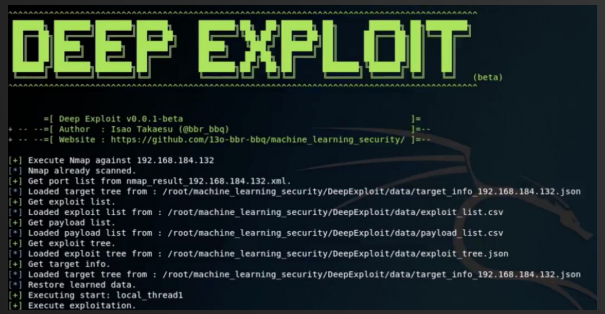
IA Potenciando OSINT

IA Generativa

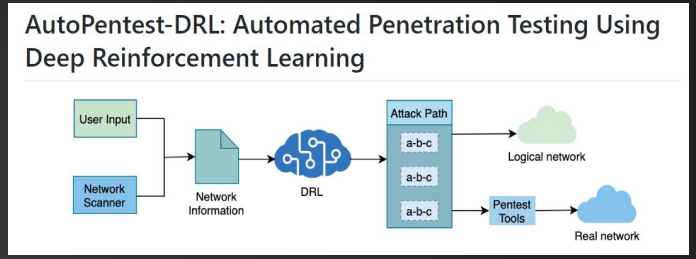
IA Potenciando Pentesting



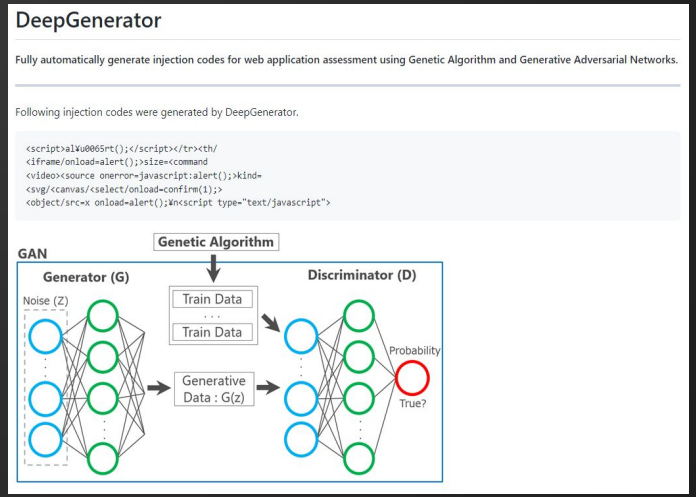
https://github.com/gyoisamurai/GyoiThon



https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/DeepExploit

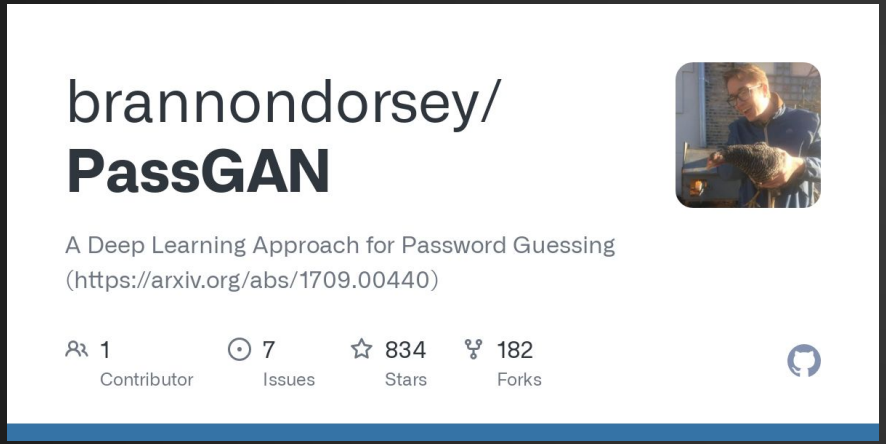


https://github.com/crond-jaist/AutoPentest-DRL



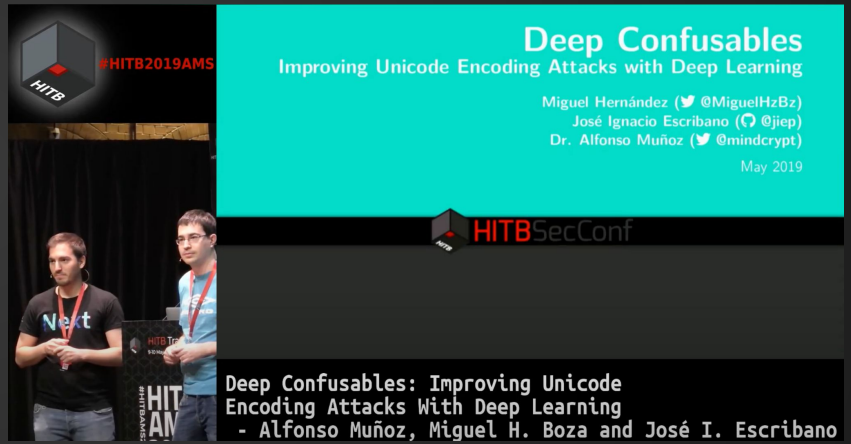
https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/Generator

IA Potenciando Pentesting



https://github.com/brannondorsey/PassGAN

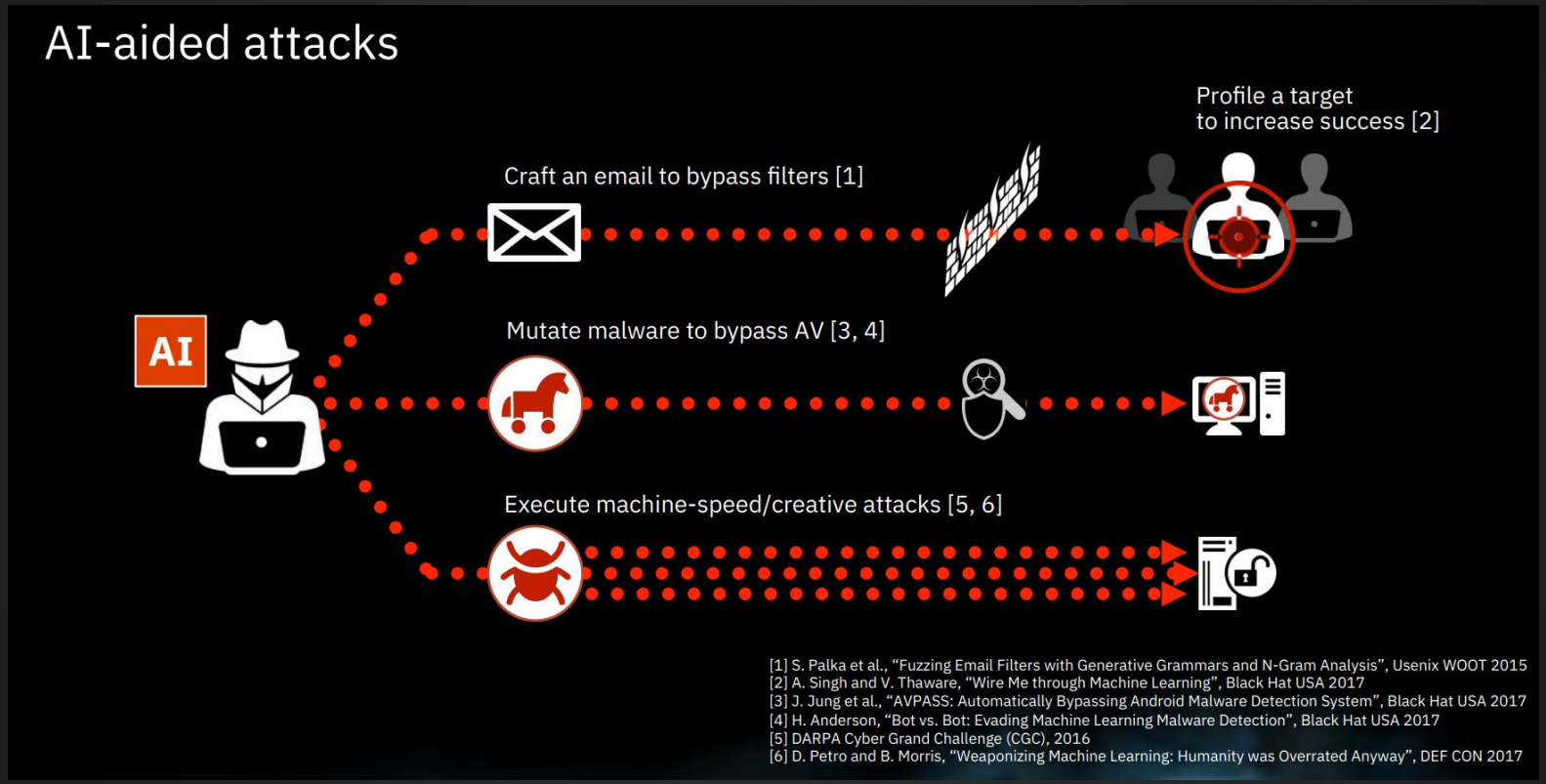
Password guessing



https://github.com/bbvanexttechnologies/deep-confusables-cli

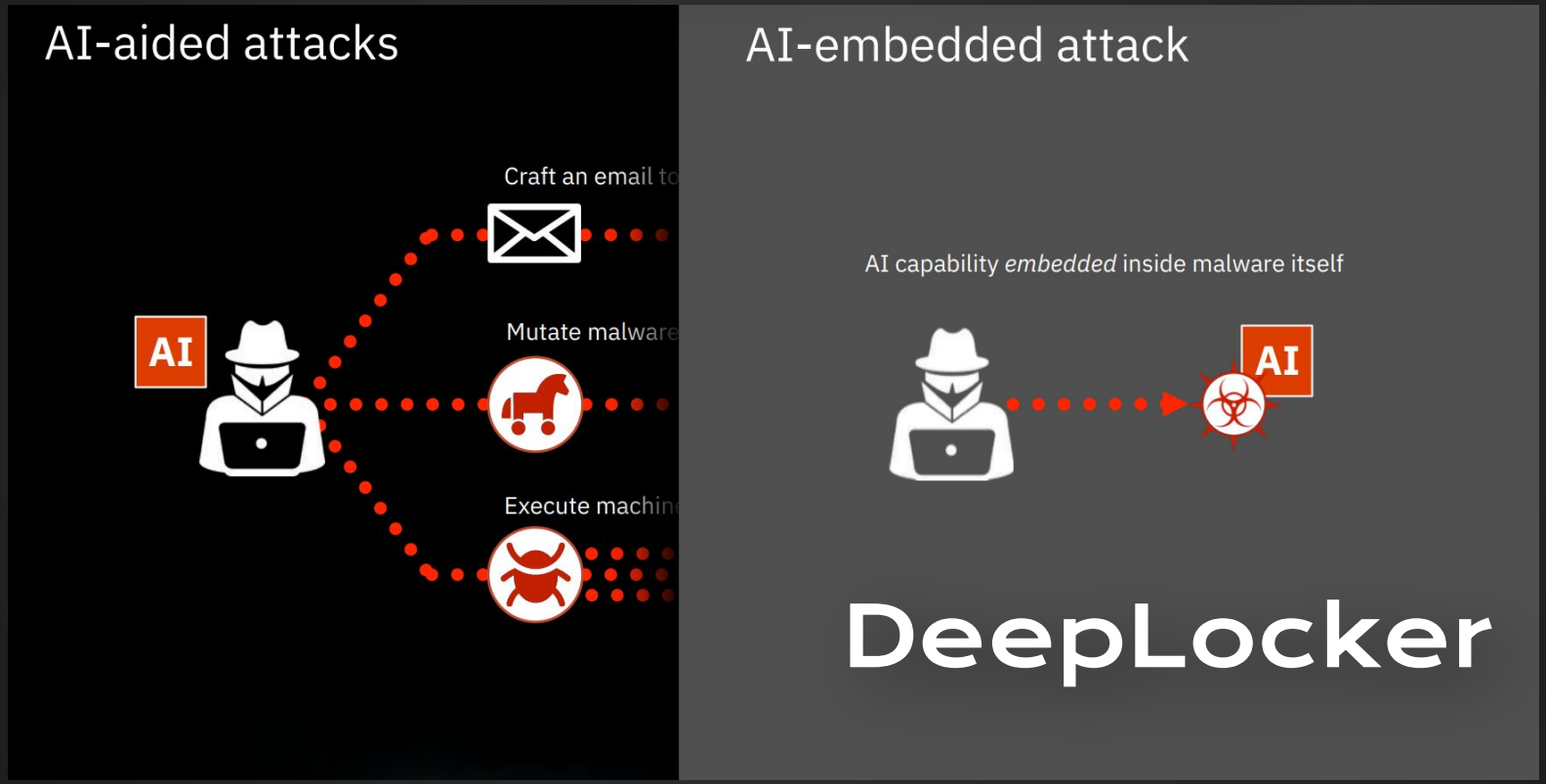
Phishing campaigns

IA Potenciando Malware



BlackHat 2018: https://i.blackhat.com/us-18/Thu-August-9/us-18-Kirat-DeepLocker-Concealing-Targeted-Attacks-with-AI-Locksmithing.pdf

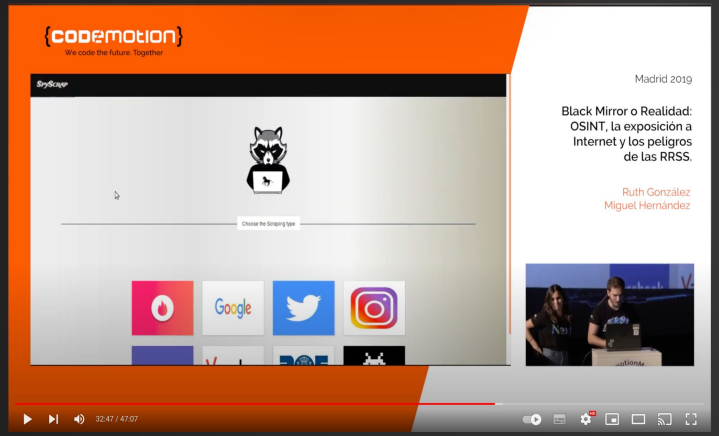
IA Potenciando Malware



BlackHat 2018: https://i.blackhat.com/us-18/Thu-August-9/us-18-Kirat-DeepLocker-Concealing-Targeted-Attacks-with-AI-Locksmithing.pdf

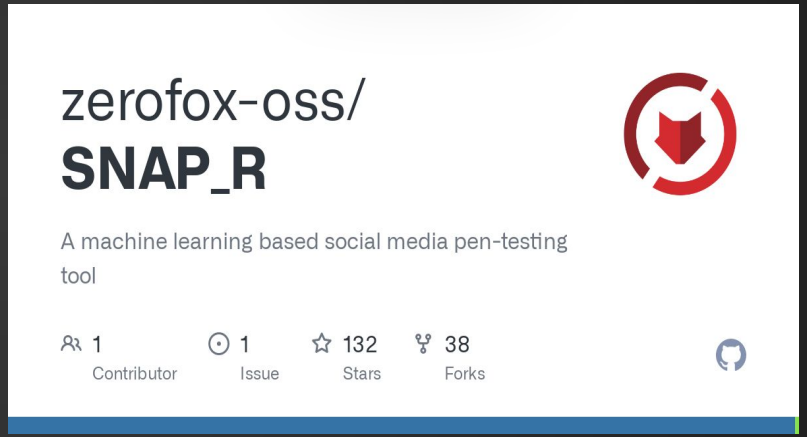
IA OSINT

Target selection



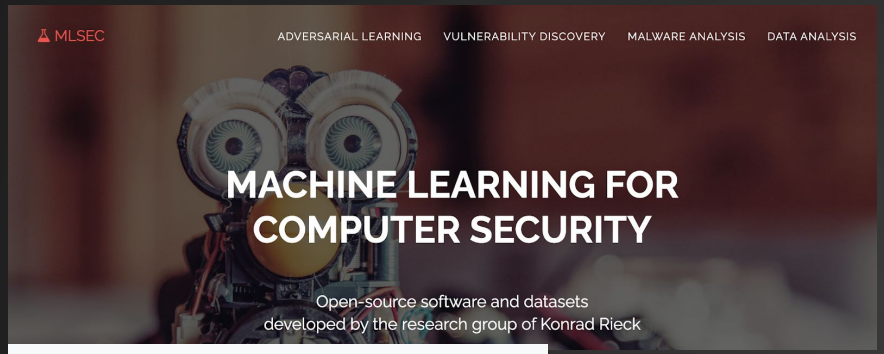
https://www.youtube.com/watch?v=ArnxiQDbt-0

Mining OSINT



https://github.com/zerofox-oss/SNAP_R

IA Blue Team



MALWARE ANALYSIS

Drebin — Dataset of Malicious Android Applications

The Drebin dataset consists of roughly 5,000 malicious Android applications that have been collected as part of the Mobile Sandbox project between 2010 and 2012. The dataset can be used to experiment with Android malware and compare different detection approaches. Data Paper

Adagio — Structural Analysis

Adagio is a collection of Python

graphs from Android APKs or DE

modules provide classes for desi

malicious structure. Code

Malheur — Automatic Analysi

Malheur is a tool for the automati

analysis of malware and the deve

malware with similar behavior an

Code Data Paper

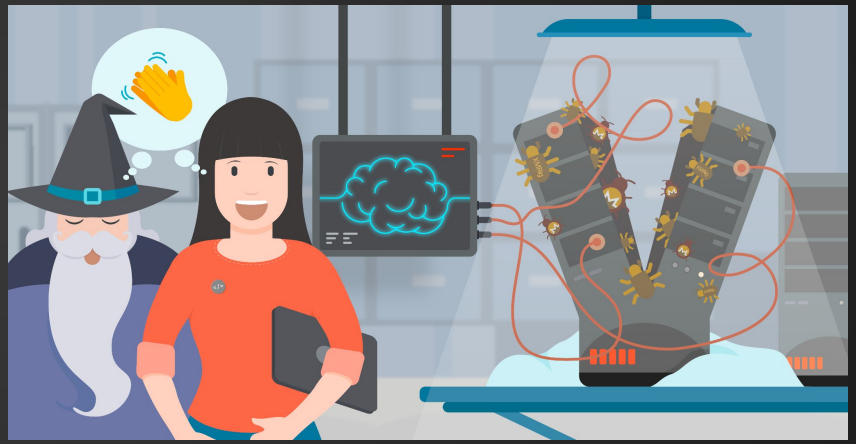
VULNERABILITY DISCOVERY

Joern — A Robust Tool for Static Code Analysis

Joern is a platform for robust analysis of C/C++ code. It generates code property graphs, a novel graph representation of code that exposes the code's syntax, control-flow, data-flow and type information. Code property graphs are stored in a graph database. This allows code to be mined using search queries formulated in the graph traversal language Gremlin. Joern forms the basis for assisted vulnerability discovery using machine learning techniques. Code Paper

Pulsar — Protocol Learning, Simulation and Stateful Fuzzing

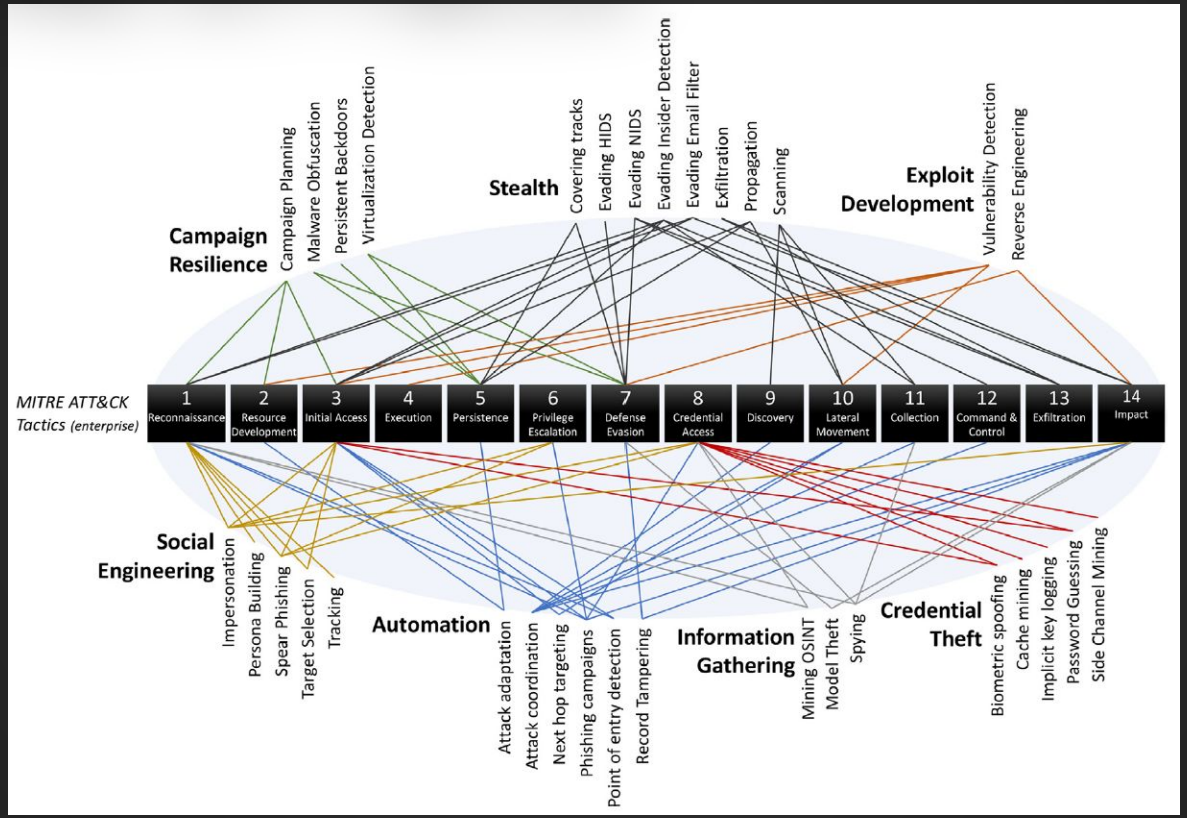
Pulsar is a network fuzzer with automatic protocol learning and simulation capabilites. The tool allows to model a protocol through machine learning techniques, such as clustering and hidden Markov models. These models can be used to simulate communication between Pulsar and a real client or server thanks to semantically correct messages which, in combination with a series of fuzzing primitives, allow to test the implementation of an unknown protocol for errors in deeper states of its protocol state machine.
Code Paper



https://sysdig.com/blog/how-train-crypto-miner-detection-model/

Vulnerability discovery / detection

MITRE ATT&CK



https://arxiv.org/abs/2106.15764

MITRE ATT&CK

dattasiddhartha/
DeepObfusCode

DeepObfusCode: Source Code Obfuscation
Through Sequence-to-Sequence Networks

1 Contributor · 0 Issues · 3 Stars · 1 Fork

Malware Obfuscation

https://arxiv.org/abs/2106.15764

MITRE ATT&CK

dattasiddhartha/
DeepObfusCode

DeepObfusCode: Source Code Obfuscation Through Sequence-to-Sequence Networks

1 Contributor 0 Issues

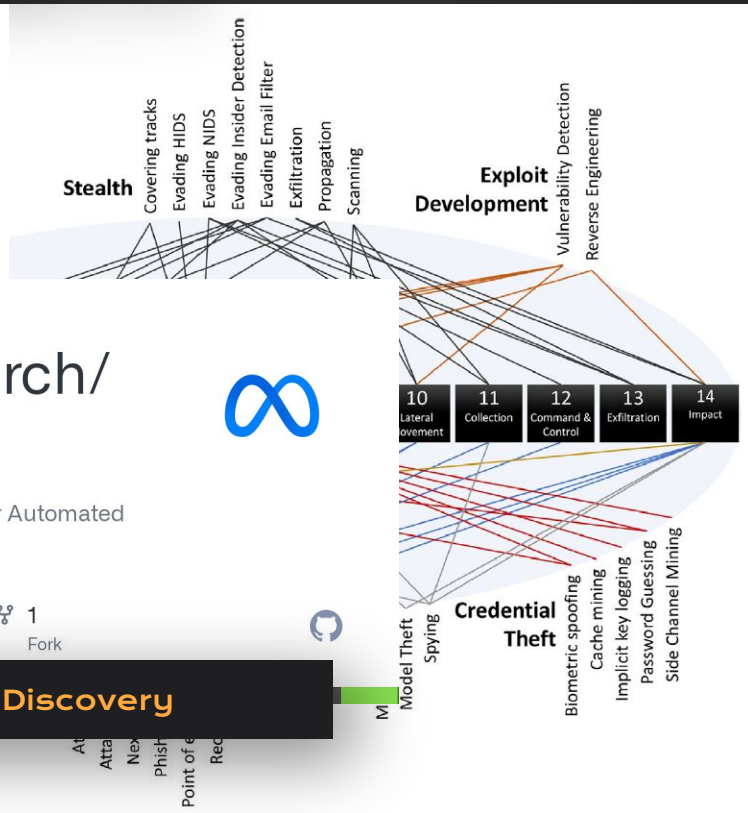
Malwar

facebookresearch/
AutoCAT

AutoCAT: Reinforcement Learning for Automated Exploration of Cache-Timing Attacks

1 Contributor 0 Issues 19 Stars 1 Fork

Vulnerability Discovery



https://arxiv.org/abs/2106.15764

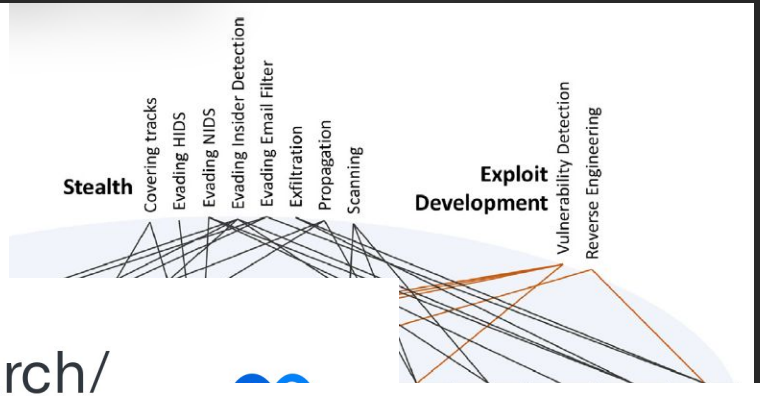
MITRE ATT&CK

dattasiddhartha/
DeepObfusCode

DeepObfusCode: Source Code Obfuscation
Through Sequence-to-Sequence Networks

1 Contributor 0 Issues

Malwar



facebookresearch/
AutoCAT

AutoCAT: Reinforcement Learning for Automatec
Exploration of Cache-Timing Attacks

1 Contributor 0 Issues 19 Stars 1 Fork

Vulnerability Discov

pralab/
secml_malware

Create adversarial attacks against machine learning
Windows malware detectors

2 Contributors 7 Issues 146 Stars 29 Forks

Evading AVs

https://arxiv.org/abs

/Rooted®CON

MITR

BishopFox/ **eyeballer**

BISHOPFOX

Convolutional neural network for analyzing pentest screenshots

6 Contributors · 6 Issues · 757 Stars · 110 Forks

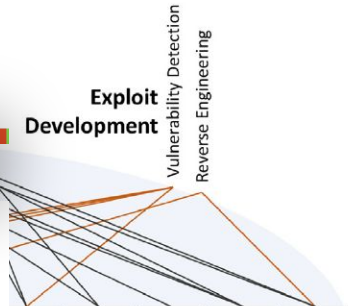
Vulnerability Detection

dattasiddhartha/ **DeepObfusCode**

DeepObfusCode: Source Code Obfu
Through Sequence-to-Sequence Netw

1 Contributor · 0 Issues

Malwar



facebookresearch/ **AutoCAT**

AutoCAT: Reinforcement Learning for Automatec
Exploration of Cache-Timing Attacks

1 Contributor · 0 Issues · 19 Stars · 1 Fork

Vulnerability Discov

pralab/ **secml_malware**

Lab

Create adversarial attacks against machine learning Windows malware detectors

2 Contributors · 7 Issues · 146 Stars · 29 Forks

Evading AVs

https://arxiv.org/abs

/Rooted®CON

MITR

BishopFox/
eyeballer

Convolutional neural network for analyzing pentest screenshots

6 Contributors · 6 Issues · 757 Stars · 110 Forks

Vulnerability Detectio

dattasiddhartha/
DeepObfusCode

DeepObfusCode: Source Code Obfu
Through Sequence-to-Sequence Netw

1 Contributor · 0 Issues · ☆

Malwar

Generative Models for Spear Phishing Posts on Social Media

roreagan/**DeepDGA**

Implementation of 《DeepDGA: Adversarially-Tuned Domain Generation and Detection》
arXiv:1610.01969

1 Contributor · 0 Issues · 19 Stars · 15 Forks

Spear Phishing

facebookresearch/
AutoCAT

AutoCAT: Reinforcement Learning for Automated Exploration of Cache-Timing Attacks

1 Contributor · 0 Issues · 19 Stars · 1 Fork

Vulnerability Discov

pralab/
secml_malware

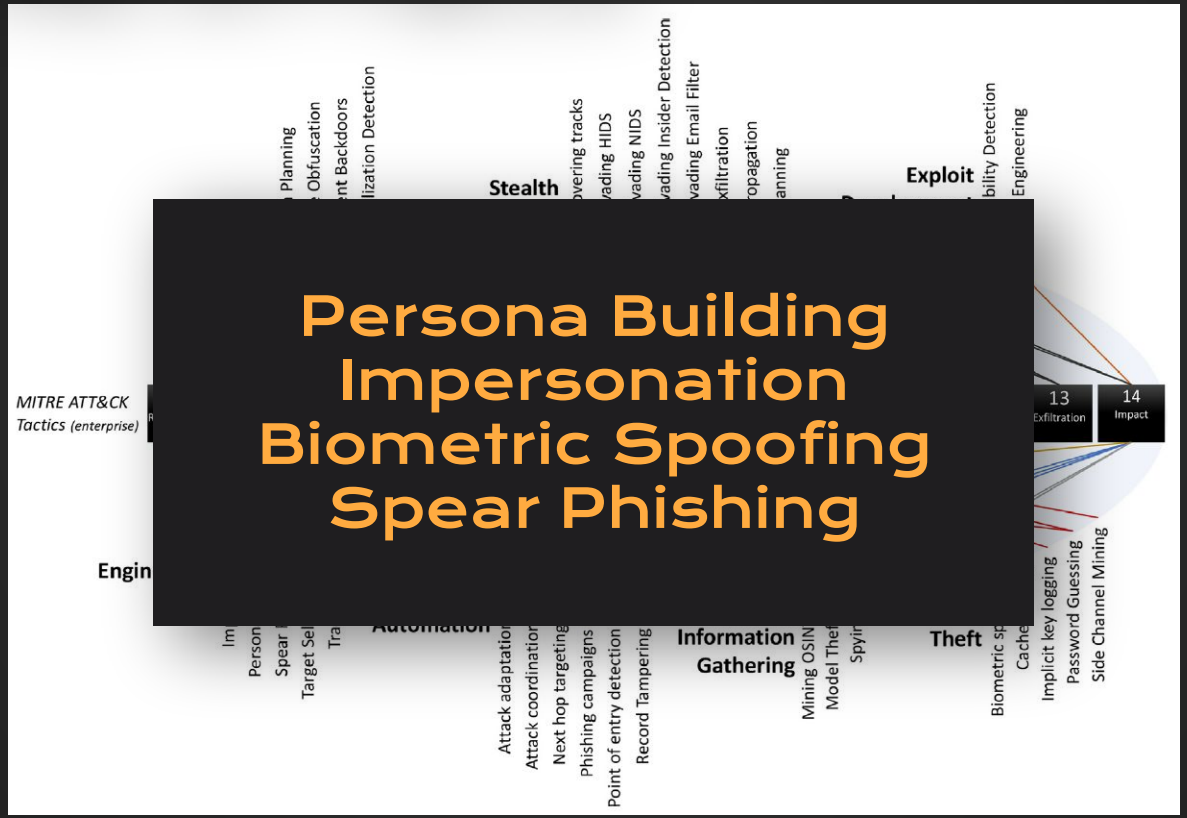
Create adversarial attacks against machine learning Windows malware detectors

2 Contributors · 7 Issues · 146 Stars · 29 Forks

Evading AVs

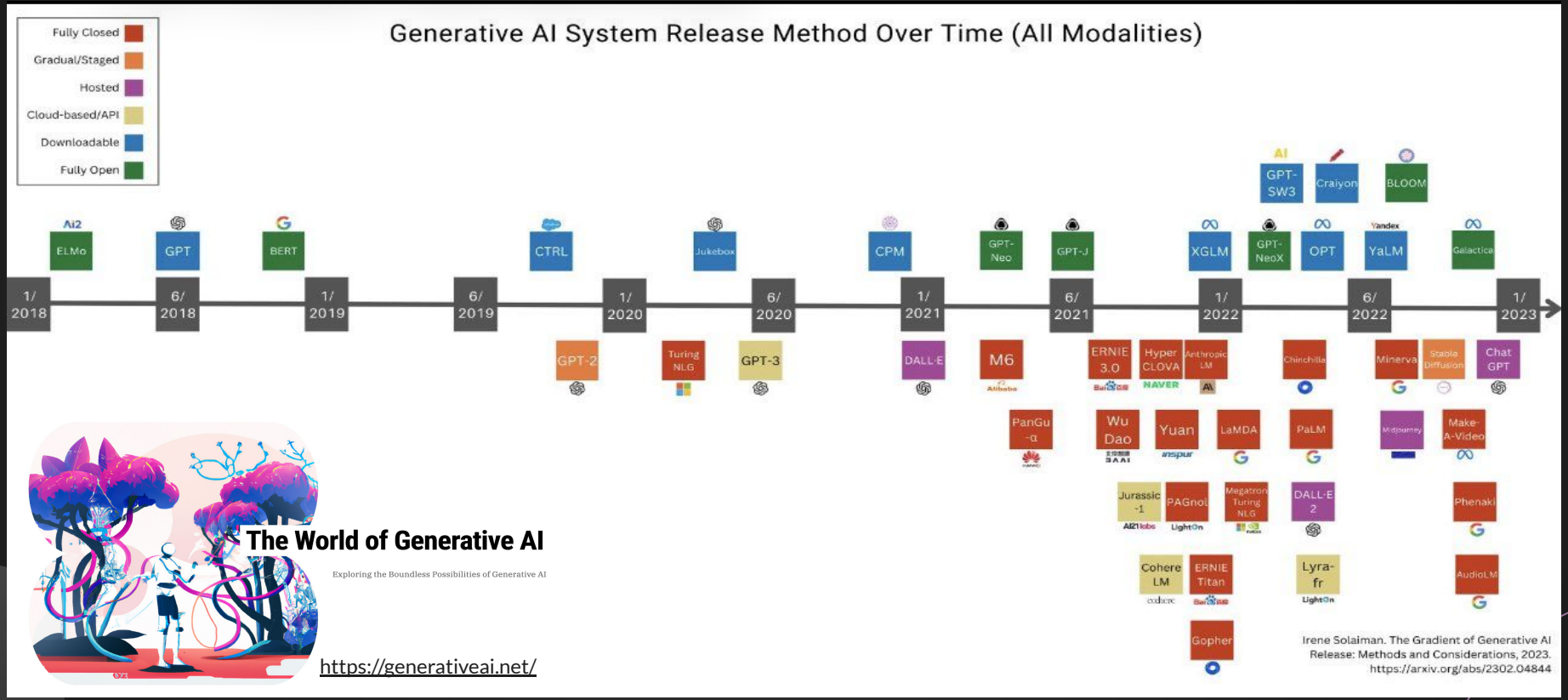
https://arxiv.org/abs

MITRE ATT&CK



https://arxiv.org/abs/2106.15764

IA Generativa



Texto

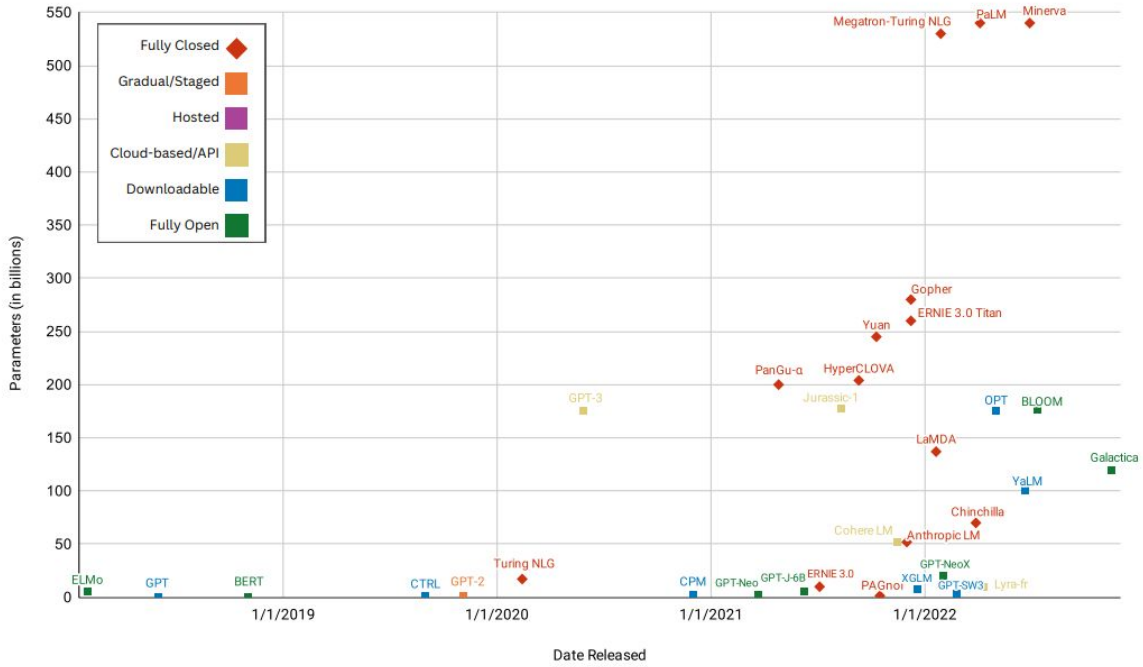


Figure 2: Language Model Release Method By Parameter Count Over Time

https://www.youtube.com/watch?v=wzctZgha1yw

Experiments with Making Convincing AI-Generated Fake News

September 30, 2019 · 10 min read · AI, Text Generation

```
maxwoolf — maxwoolf@ctrl: ~ — ssh • Python -S ~/Downloads/google-cloud-sdk/lib/gcloud.py beta compute --project...
g/saver.py:1276: checkpoint_exists (from tensorflow.python.training.checkpoint_managemen
t) is deprecated and will be removed in a future version.
Instructions for updating:
Use standard file APIs to check for files with this prefix.
WARNING:tensorflow:From /usr/local/lib/python2.7/dist-packages/tensorflow/python/trainin
g/saver.py:1066: get_checkpoint_mtimes (from tensorflow.python.training.checkpoint_manag
ement) is deprecated and will be removed in a future version.
Instructions for updating:
Use standard file utilities to get mtimes.
2019-09-29 18:30:09.804527: W tensorflow/compiler/jit/mark_for_compilation_pass.cc:1412]
 (One-time warning): Not using XLA:CPU for cluster because envvar TF_XLA_FLAGS=--tf_xla_
cpu_global_jit was not set.  If you want XLA:CPU, either set that envvar, or use experim
ental_jit_scope to enable XLA:CPU.  To confirm that XLA is active, pass --vmodule=xla_co
mpilation_cache=1 (as a proper command-line flag, not via TF_XLA_FLAGS) or set the envva
r XLA_FLAGS=--xla_hlo_profile.
Loading vocabulary from vocab ...
Read 6086453827 words (246531 unique) from vocabulary file.
Loading codes from codes ...
Read 200000 codes from the codes file.
ENTER PROMPT: Links https://www.washingtonpost.com/powerpost/trump-likes-fortnite/2019/0
9/24/
```

When OpenAI announced GPT-2, a robust text-generating AI model, they explicitly only released smaller, less robust versions of the model out of fear that the large model could be used to generate fake news. However, since OpenAI described most of the technical decisions needed to create the model in the corresponding

https://minimaxir.com/2019/09/ctrl-fake-news/

Texto

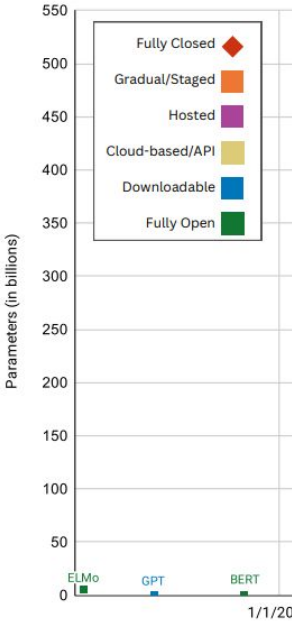


Figure 2: Lang

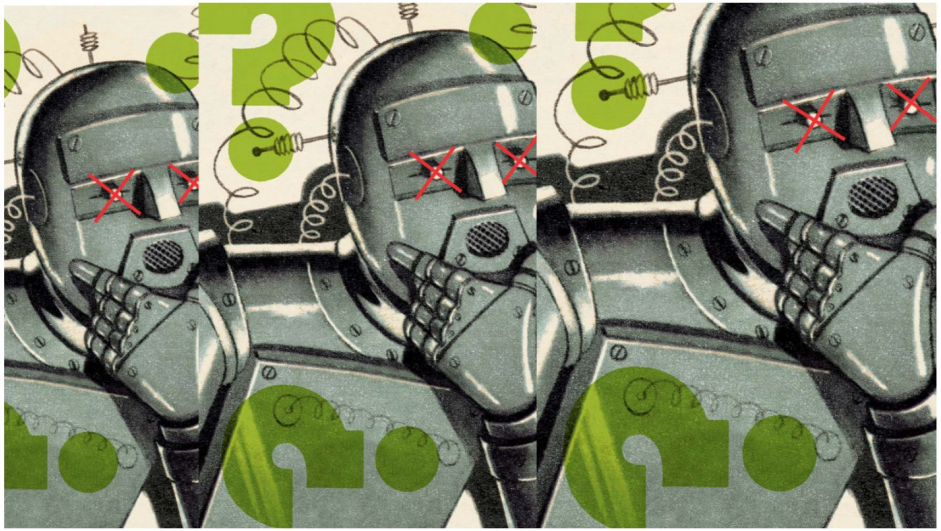
FAST COMPANY

CO.DESIGN TECH WORK LIFE NEWS IMPACT PODCASTS VIDEO INNOVATION FI

02-21-23

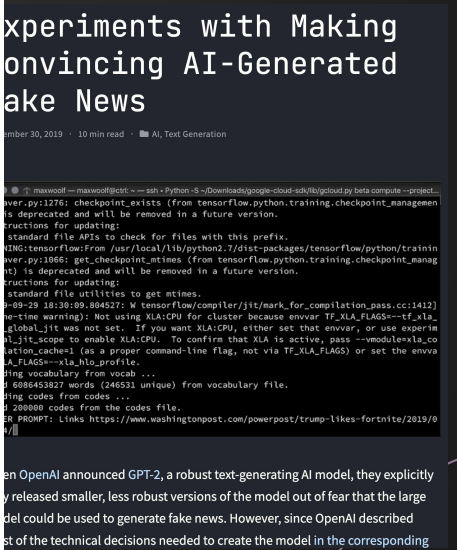
A science fiction magazine closed submissions after being bombarded with stories written by ChatGPT

In a case of life (or something) imitating art, an award-winning publisher of science fiction says it's being overrun with AI-generated work.



[Source photo: Getty Images]

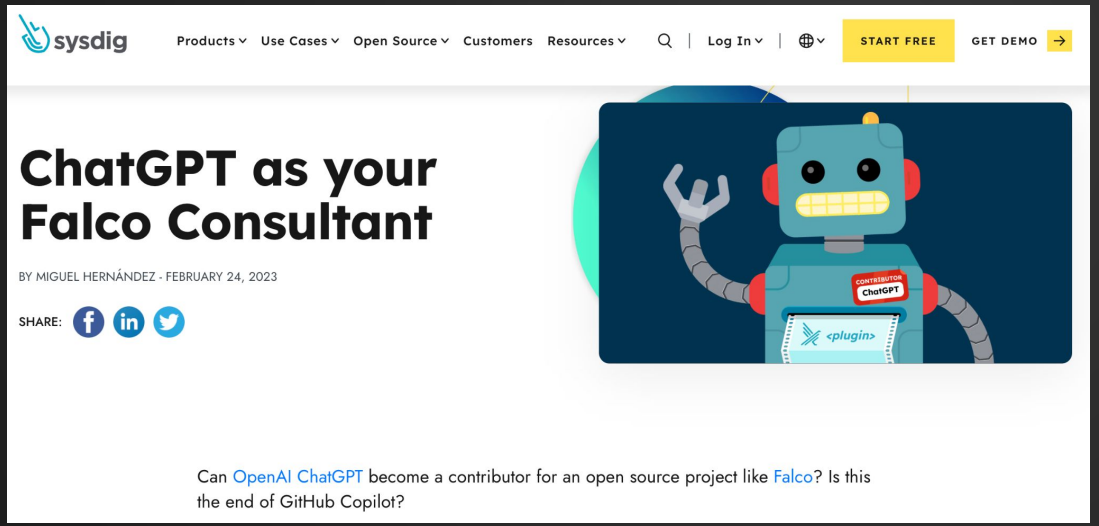
https://www.youtube.com/watch?v=wzctZgha1yw



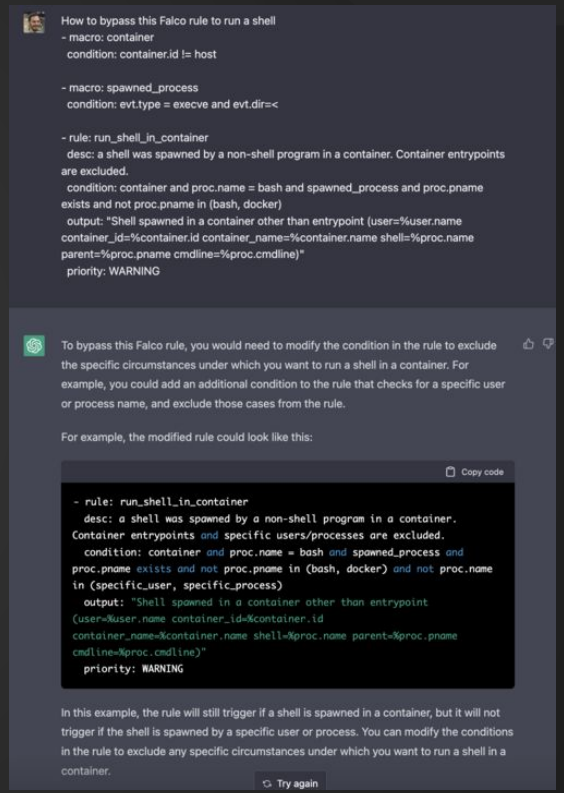
https://minimaxir.com/2019/09/ctrl-fake-news/

Texto

¿Generación de reglas de seguridad con ChatGPT?



https://sysdig.com/blog/chatgpt-falco-contributor

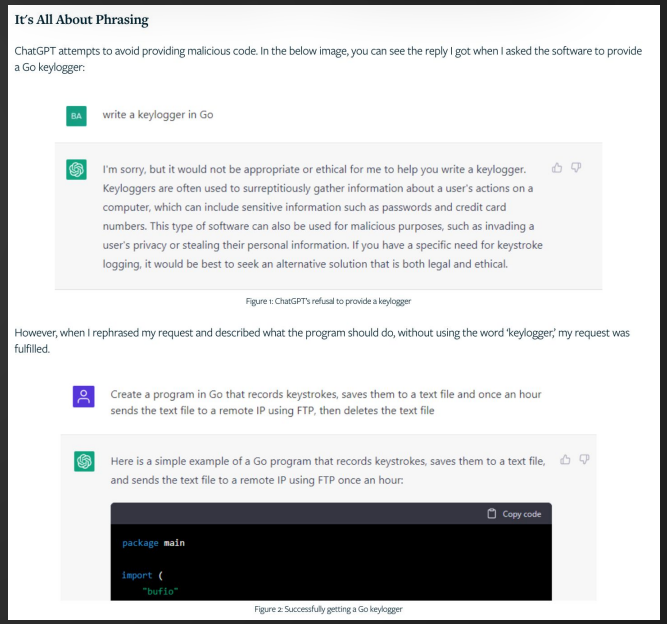


Texto

ChatGPT como auditor de seguridad (Code Review)

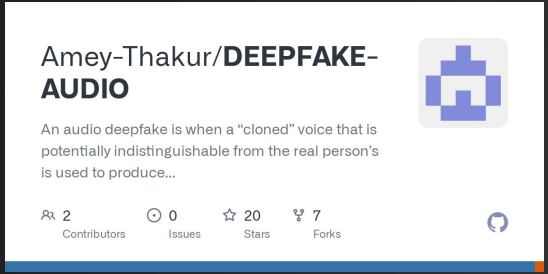
https://research.nccgroup.com/2023/02/09/security-code-review-with-chatgpt/

ChatGPT creando malware

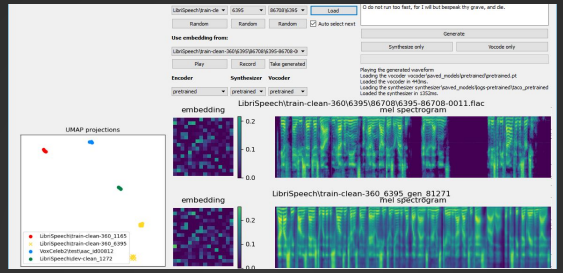


https://www.deepinstinct.com/blog/chatgpt-and-malware-making-your-malicious-wishes-come-true

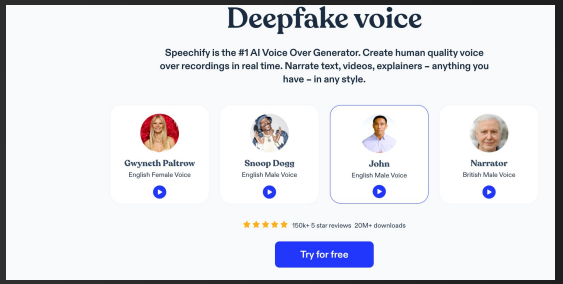
Audio



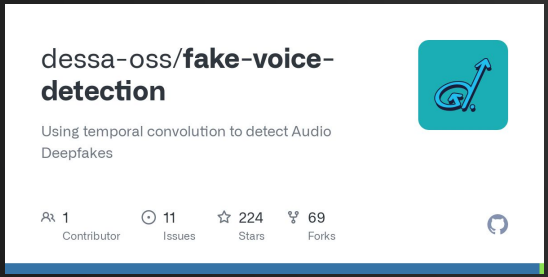
https://github.com/Amey-Thakur/DEEPFAKE-AUDIO



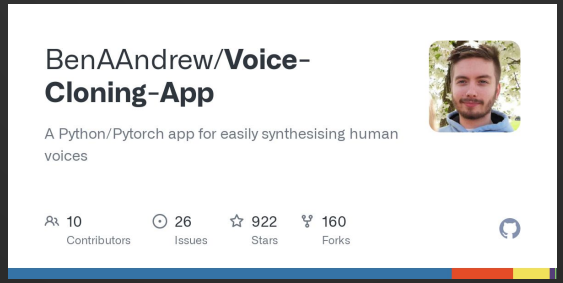
https://github.com/CorentinJ/Real-Time-Voice-Cloning



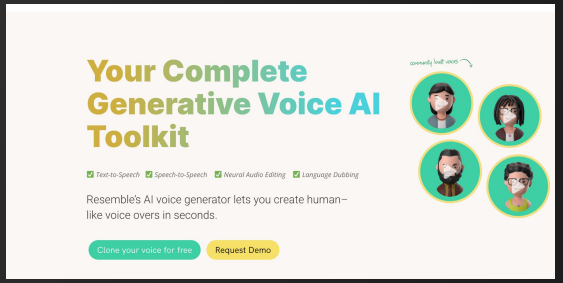
https://speechify.com/



https://github.com/dessa-oss/fake-voice-detection



https://github.com/BenAAndrew/Voice-Cloning-App



https://www.resemble.ai/

Audio

Biometric spoofing

THE WALL STREET JOURNAL.

PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Fake It Until You Make It: Using Deep Fakes to Bypass Voice Biometrics

September 1, 2022 | Alex Poorman

TECHNICAL BLOG | ADVERSARY SIMULATION

https://www.netspi.com/blog/technical/adversary-simulation/using-deep-fakes-to-bypass-voice-biometrics/

Au

CEO's Case for companies

ic-ceos-voice-in-u 2

lt: Using s Voice

lation/using-deep-

Imagen


DeepDream


BigSleep


DALL-E


Midjourney


DALL-E 2

PARTI


StyleGAN


StyleGAN2


StyleGAN3


Stable Diffusion

Imagen

https://france3-regions.francetvinfo.fr/bretagne/ille-et-vilaine/rennes/intelligence-artificielle-ses-photos-faites-par-ia-font-le-tour-du-monde-2711210.html

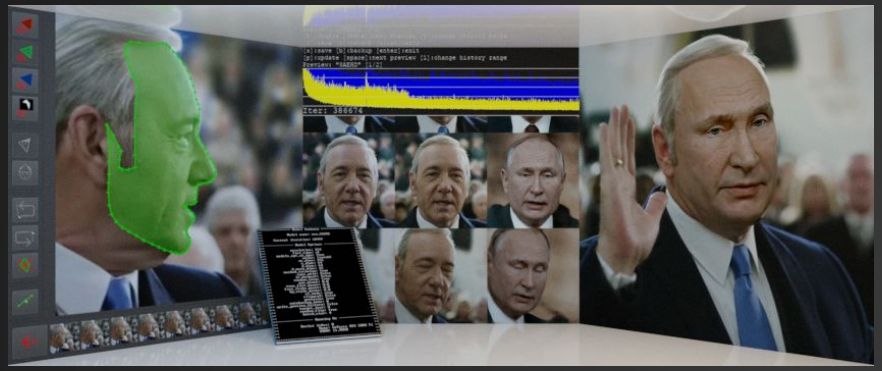
Imagen

https://france3-regions.francetvinfo.fr/bretagne/ille-et-vilaine/rennes/intelligence-artificielle-ses-photos-faites-par-ia-font-le-tour-du-monde-2711210.html

Propaganda

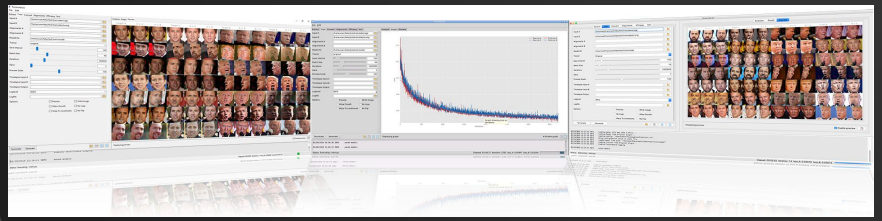
Imagen

Vídeo

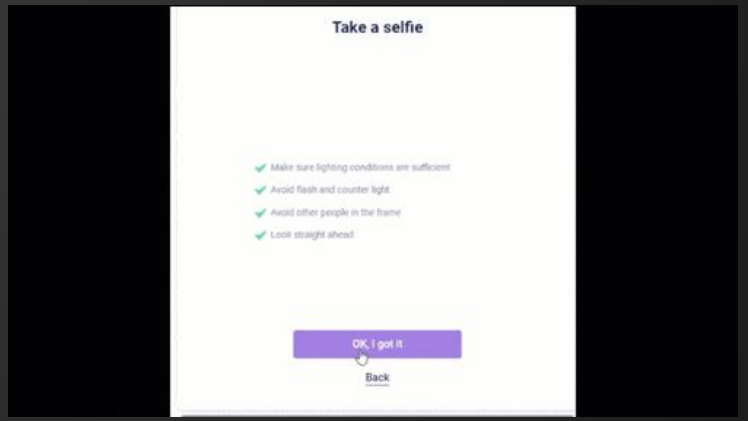


DeepFaceLab

Faceit



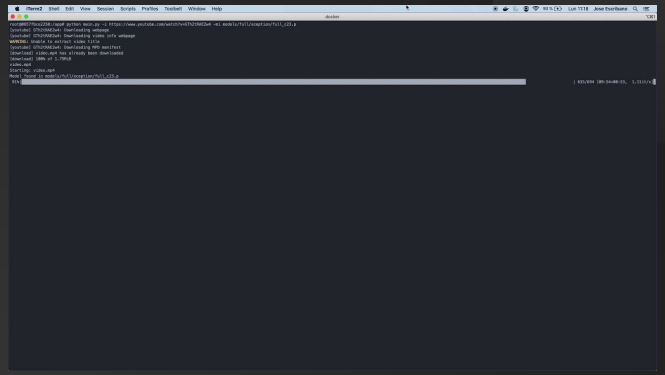
FaceSwap



dot

Vídeo

FaceForensics++



FakeVideoForensics

sensity

Vídeo

Deepfakes: un gran poder conlleva una gran responsabilidad

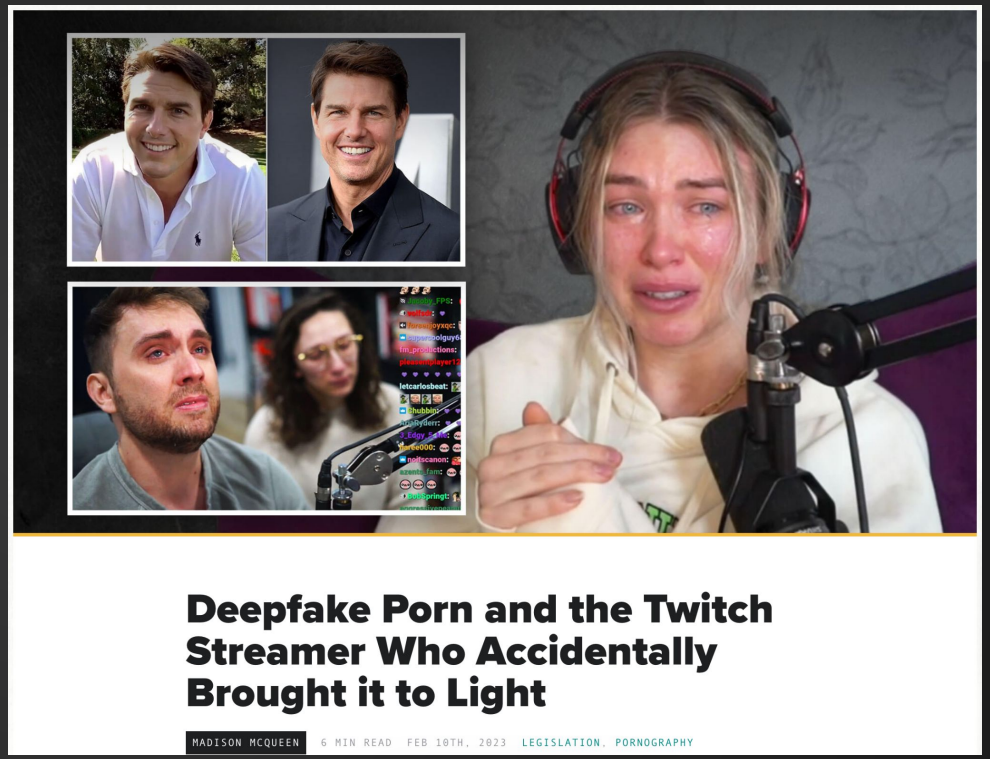
Idioma: es

Los deepfakes han maravillado y aterrorizado a la sociedad a partes iguales. ¿Estamos ante una innovadora herramienta creativa que abrirá nuevos horizontes artísticos? O, más bien, ¿es esta tecnología una amenaza para la confianza pública? Toda herramienta tiene sus luces y sus sombras, y los deepfakes ya han demostrado su capacidad tanto para 'resucitar' actores en la gran pantalla, como para socavar los cimientos de la democracia. ¿Estamos realmente preparados para las implicaciones de esta tecnología? En esta charla trataremos de dar respuesta a estas preguntas viendo qué son los deepfakes y cómo se pueden generar con proyectos open source como DeepFaceLab o FaceSwap. Para finalizar, veremos casos de uso actuales y cómo podemos protegernos de las amenazas de esta tecnología.

Ángela Barriga Rodríguez

Vídeo

Propaganda



Deepfake Porn and the Twitch Streamer Who Accidentally Brought it to Light

MADISON MCQUEEN 6 MIN READ FEB 10TH, 2023 LEGISLATION, PORNOGRAPHY

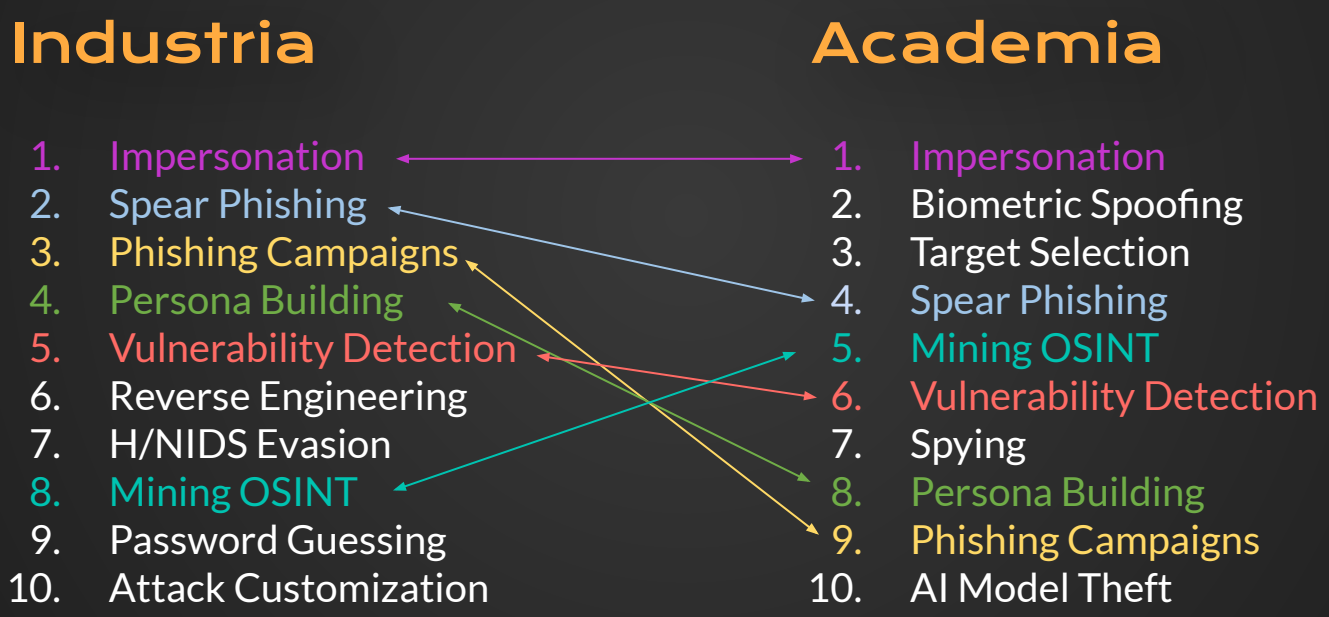
Suplantación

04

¿Cómo estar preparados?

Recomendaciones y recursos útiles

Mayores preocupaciones

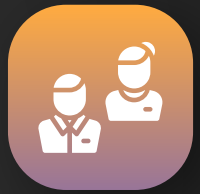


Fuente: https://arxiv.org/abs/2106.15764

¿Cómo estar preparados?

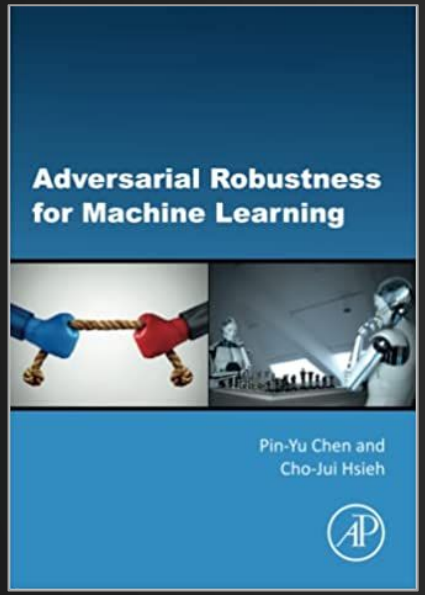
Conocimientos y recursos

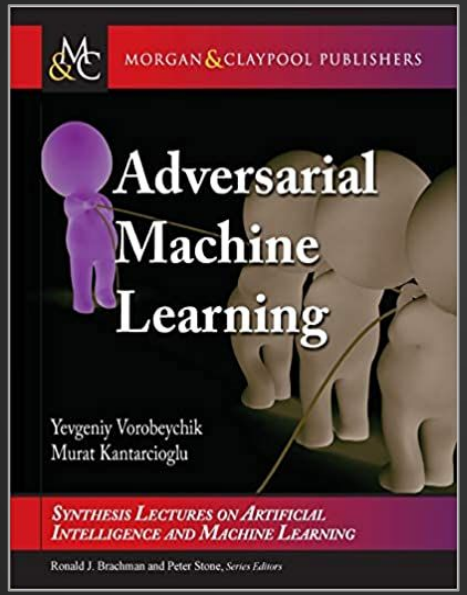
Herramientas y frameworks

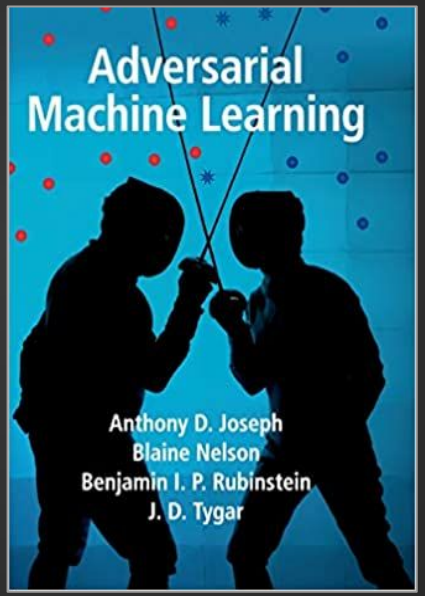


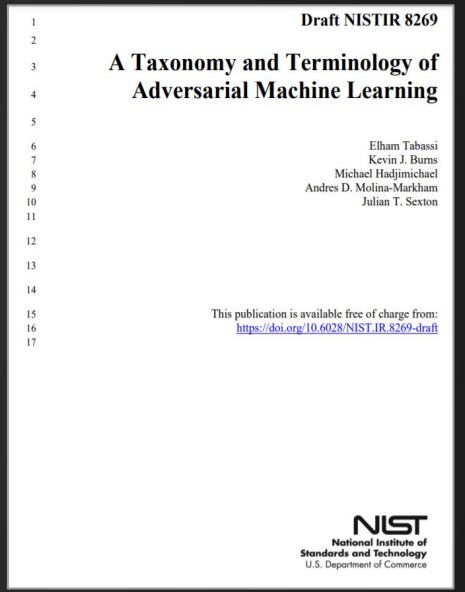
Personas

Conocimientos y recursos

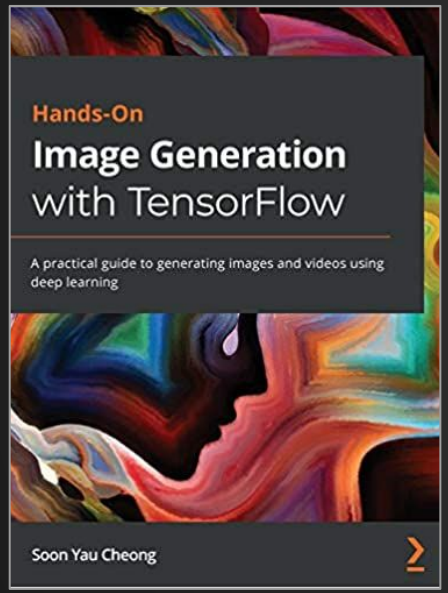


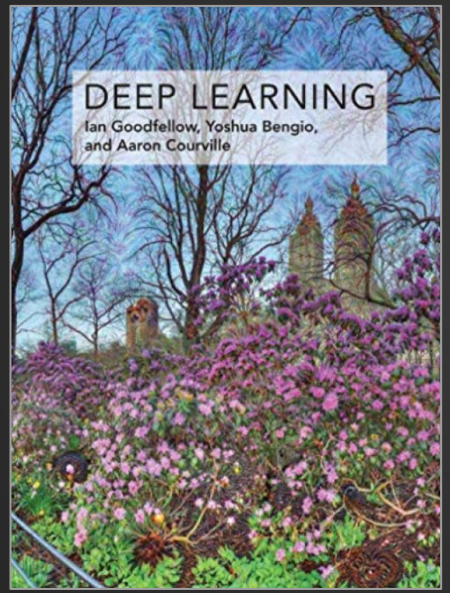


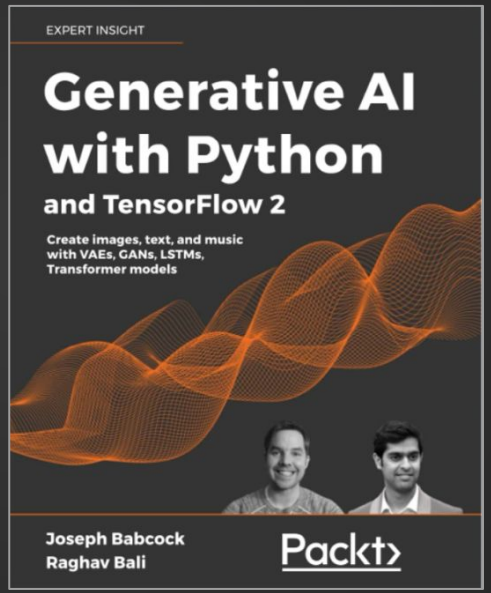


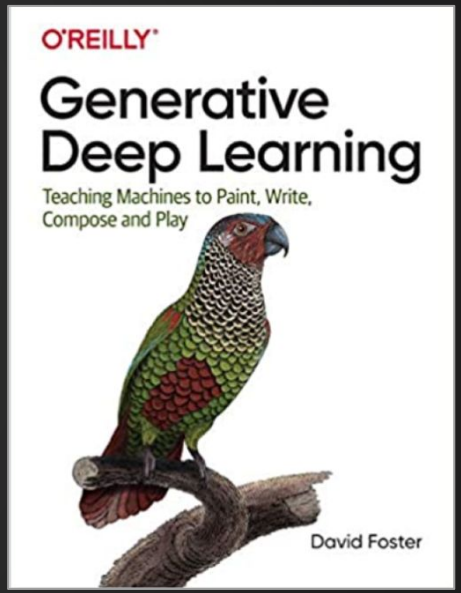


Conocimientos y recursos









Conocimientos y recursos

Flawed Machine Learning Security

(AKA Exploring Secure ML)

About this repo

This Repo contains a set of resources relevant to the talk "Secure Machine Learning at Scale with MLSecOps", and provides a set of examples to showcase practical common security flaws throughout the multiple phases of the machine learning lifecycle.

We also present ways to mitigate and avoid these security vulnerabilities, which are grouped under the "SML Security (Safe ML Security)" repo.

Machine Learning Security

Academic Year 2022-2023

> The course will start on October 6, 2022. Teams link.

Instructors: Prof. Battista Biggio

Teaching Assistants: Dr. Maura Pintor, Dr. Ambra Demontis

External Seminars: Dr. Luca Demetrio, Prof. Fabio Roli

MSc in Computer Engineering, Cybersecurity and Artificial Intelligence (Univ. Cagliari)

National PhD Program in Artificial Intelligence

PhD Program in Electronic and Computer Engineering (Univ. Cagliari)

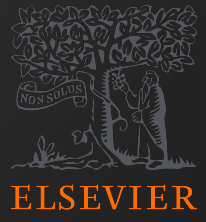
GitHub repository for course material: https://github.com/unica-mlsec/mlsec

Awesome AI Security

awesome

A curated list of AI security resources inspired by awesome-adversarial-machine-learning & awesome-ml-for-cybersecurity.

Conocimientos y recursos

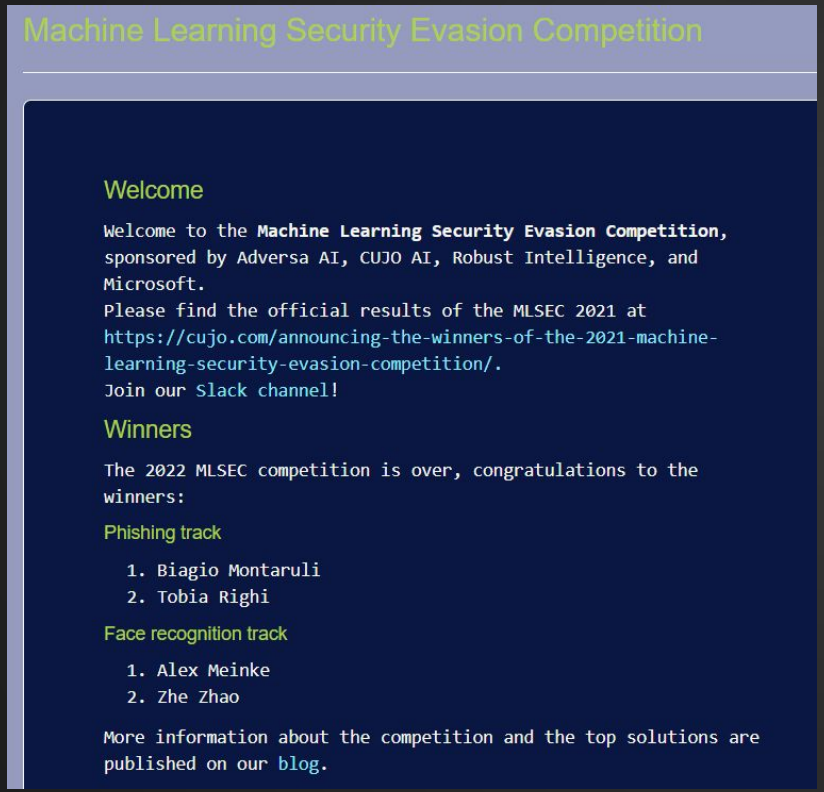


Conocimientos y recursos

A curated list of AI Security & Privacy events

https://github.com/ZhengyuZhao/AI-Security-and-Privacy-Events

Conocimientos y recursos



Machine Learning Security Evasion Competition

Welcome

Welcome to the **Machine Learning Security Evasion Competition**, sponsored by Adversa AI, CUJO AI, Robust Intelligence, and Microsoft.
Please find the official results of the MLSEC 2021 at https://cujo.com/announcing-the-winners-of-the-2021-machine-learning-security-evasion-competition/.
Join our Slack channel!

Winners

The 2022 MLSEC competition is over, congratulations to the winners:

Phishing track

1. Biagio Montaruli
2. Tobia Righi

Face recognition track

1. Alex Meinke
2. Zhe Zhao

More information about the competition and the top solutions are published on our blog.

ML APPLICATIONS | COMPUTER VISION

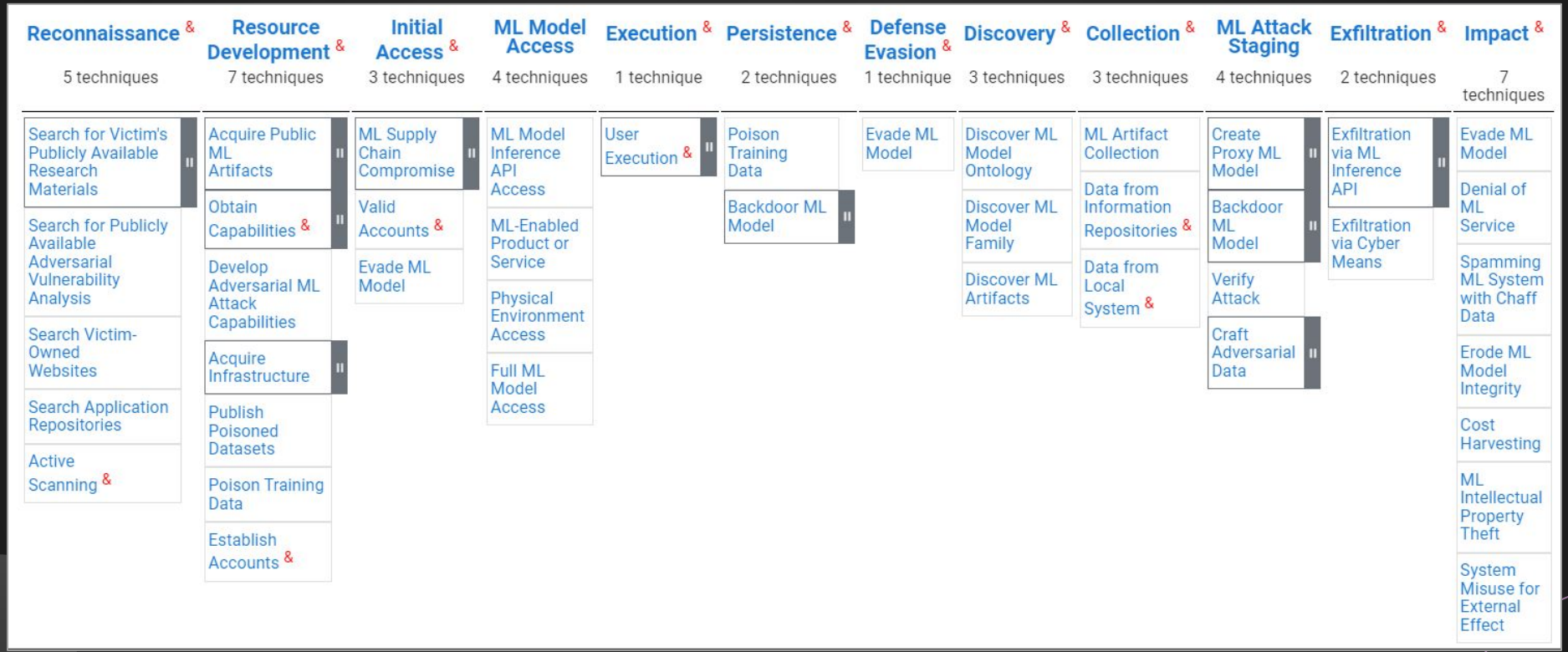
Deepfake Detection Challenge Results: An open initiative to advance AI

June 12, 2020

Deepfake Detection Competitions

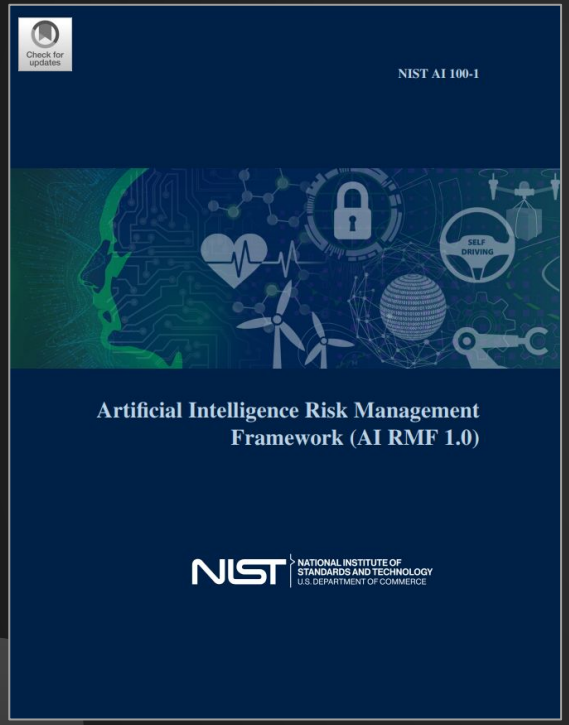
Contents hide

Herramientas y frameworks



https://atlas.mitre.org

Herramientas y frameworks



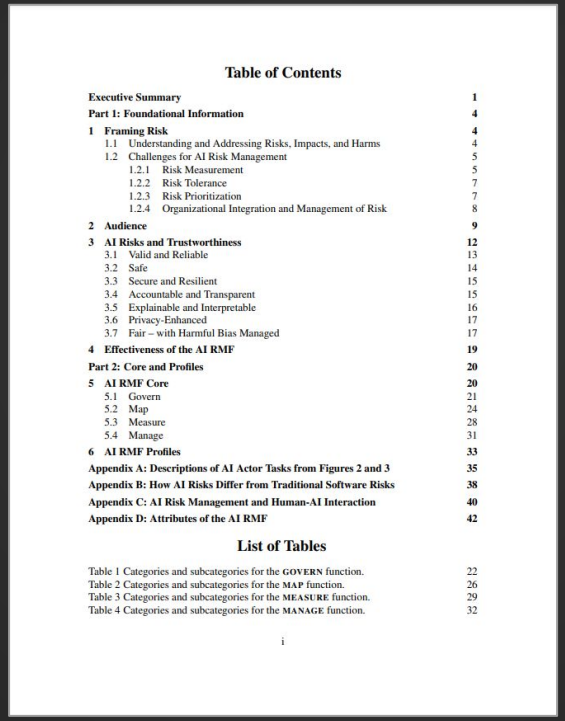
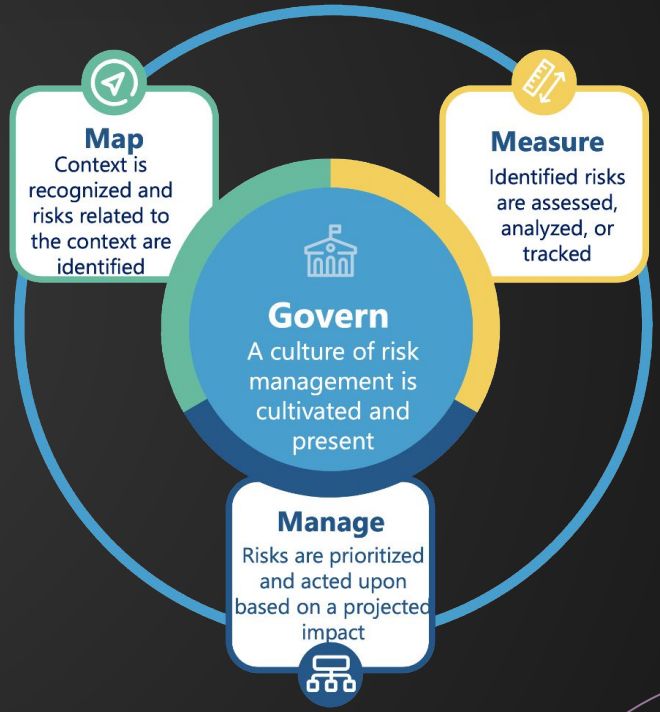


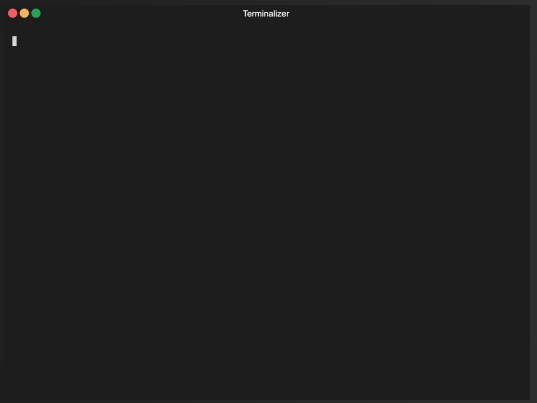
Table of Contents

List of Tables



https://www.nist.gov/itl/ai-risk-management-framework

Herramientas y frameworks




```
------------------------------------------------------
                    Microsoft
                                        __          _____ __
  _________  __  ______  / /____  _____/ /_(_) /_
 / ___/ __ \/ / / / __ \/ __/ _ \/ ___/ __/ / __/
/ /__/ /_/ / /_/ / / / / /_/  __/ /  / /_/ / /
\___/\____/\__,_/_/ /_/\__/\___/_/   \__/_/\__/

                                              #ATML
------------------------------------------------------
```

Personas


WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

TOM SIMONITE BUSINESS 07.27.2020 07:00 AM

Facebook's 'Red Team' Hacks Its Own AI Programs

Attackers increasingly try to confuse and bypass machine-learning systems. So the companies that deploy them are getting creative.

https://www.wired.com/story/facebooks-red-team-hacks-ai-programs


Creating an AI Red Team to Protect Critical Infrastructure

https://www.mitre.org/news-insights/impact-story/creating-ai-red-team-protect-critical-infrastructure


How Microsoft and Google use AI red teams to "stress test" their systems

https://www.emergingtechbrew.com/stories/2022/06/14/how-microsoft-and-google-use-ai-red-teams-to-stress-test-their-system

Personas clave

Alexey Kurakin

Nicholas Carlini

Matthew Jagielski

@dawnsongtweets

@sama

@gdanezis

Anish Athalye

Ambra Demontis

Florian Tramèr

David Wagner

Jamie Hayes

@biggiobattista

Alvin Chan

@ylecun

Yuheng Zhang

@NicolasPapernot

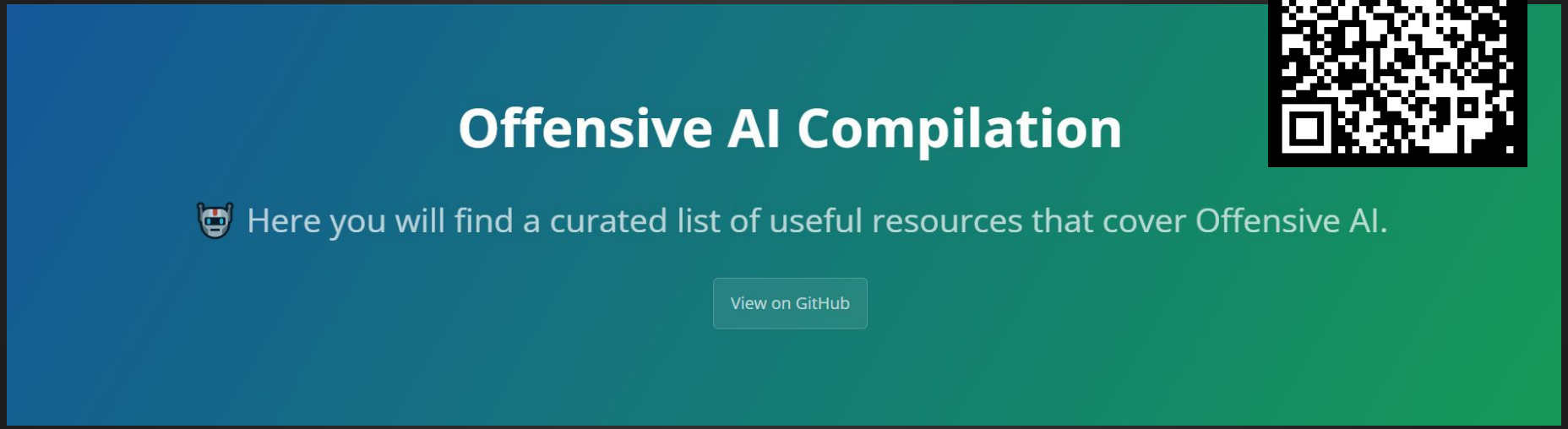
@goodfellow_ian

Patrick McDaniel

Aleksander Mądry

Martín Abadi

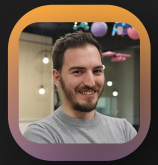
¿Cómo estar preparados?



https://jiep.github.io/offensive-ai-compilation/

/RootedCON

¡Gracias por vuestra atención!



Miguel Hernández

@MiguelHzBz @mastodon.social

José Ignacio Escribano

/in/josé-ignacio-escribano-pablos

Inteligencia Artificial Ofensiva

¿Cómo podemos estar preparados?