Hacker vs AI

Perspectives from an ex-spy

MILEVA
SECURITY LABS

Mileva SECURITY LABS

WHO AM I



Mileva Security Labs

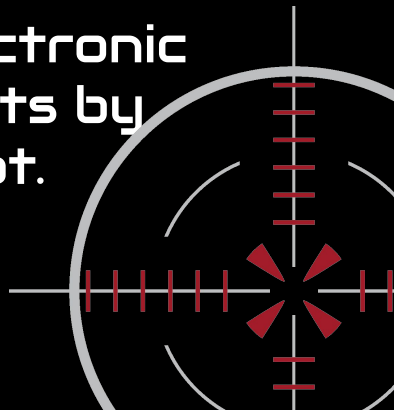
@HarrietHacks

MILEVA SECURITY LABS

ASD AUSTRALIAN SIGNALS DIRECTORATE

The Australian Government intelligence agency responsible for foreign signals intelligence and cyber security.

ASD collects and analyses electronic signals from communications systems, radio frequencies and electronic data transmissions. They defend Australia's interests by mastering technology to inform, protect and disrupt.



Mileva Security Labs

MY RESEARCH



THE RACE TO ROBUSTNESS: EXPLOITING FRAGILE MODELS FOR URBAN CAMOUFLAGE AND THE IMPERATIVE FOR MACHINE LEARNING SECURITY

Harriet Farlow
School of Engineering and Information Technology
University of New South Wales
Canberra

Matthew Garratt
School of Engineering and Information Technology
University of New South Wales
Canberra

Gavin Mount
School of Humanities and Social Sciences
University of New South Wales
Canberra

Tim Lynar
School of Engineering and Information Technology
University of New South Wales
Canberra

26 Jun 2023

Deceive, Disrupt, Disclose: The 3D Model for AI Security

Harriet Farlow[1,2][0000−0003−0822−1628], Tania Sadhani[2], Matthew Garratt[1][0000−0003−0222−430X], Gavin Mount[1][0000−0002−0956−8327], and Timothy Lynar[1][0000−0001−7934−5658]

[1] School of Engineering and Information Technology University of New South Wales, Canberra, Australia
[2] Mileva Security Labs

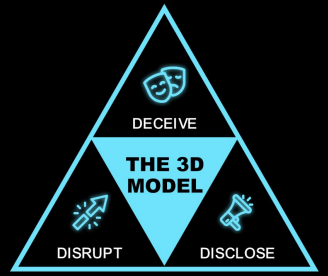
Abstract. The evolving Artificial Intelligence (AI) landscape necessitates a maturation of the frameworks, processes and policies for visibility and management of AI-related risk. Adversarial machine learning (AML) is the academic field that represents the offensive side of AI security, and provides methodologies for AI security risks. While frameworks and guidelines currently exist for AI safety and cyber security risk, there exists a gap in the practical management of AI security risks. This paper therefore conducts a systematic analysis of AML attacks and defenses and places these techniques in the context of existing AI risk frameworks. We present a novel threat-centric security model for AI Security, the 3D Model, which maps those AML techniques to three axes - Disruption,

Is AI going through a securitization process: a review of public AI policy informed by security studies

Harriet Farlow (0000-0003-0822-1628), Matthew Garratt (0000-0003-0222-430X), Gavin Mount (0000-0002-0956-8327), Timothy Lynar (0000-0001-7934-5658)

School of Systems and Computing, University of New South Wales, Canberra, Australia

Artificial Intelligence (AI) systems are undergoing rapid adoption, with the promise of increased efficiency and decreased costs for organizations across all industries. However the addition of new AI systems also introduces new risks, prompting the creation of AI policies to address this. The Organization for Economic Cooperation and Development (OECD) maintains a database of 1937 public AI policies from 69 countries. These policies frequently reference AI





MILEVA SECURITY LABS

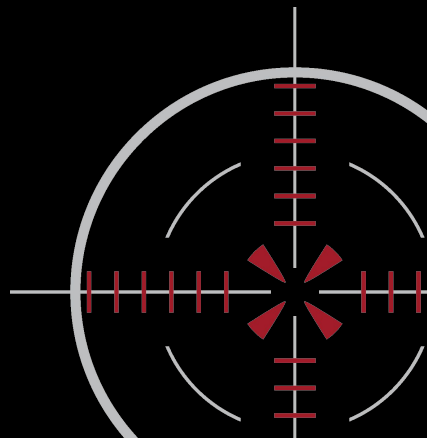
OUR OBJECTIVES



SECURITY THEORY

AI FOR CYBER DEFENSE AND OFFENSE

NATIONAL SECURITY IMPLICATIONS



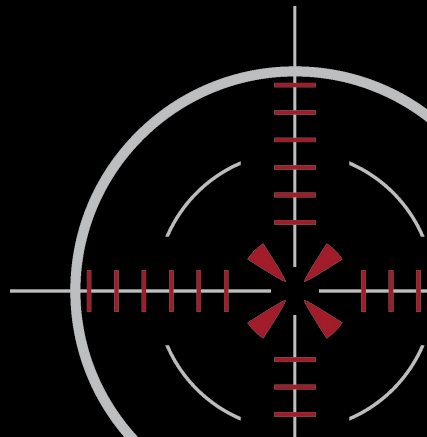
MILEVA SECURITY LABS

BUT FIRST

SAY HI TO THE PERSON NEXT TO YOU

EXPLAIN WHY YOU CAME TO THIS TALK

TELL YOUR FAVOURITE SCI-FI MOVIE



MILEVA SECURITY LABS

FAVOURITE SCI-FI?







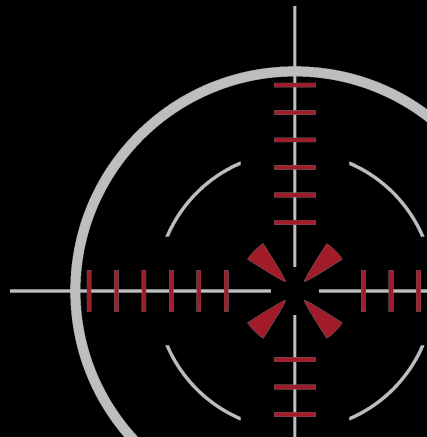
MILEVA SECURITY LABS

SECURITY STUDIES

REFERENT OBJECT

DIRECTION OF PROTECTION

SECURITIZATION

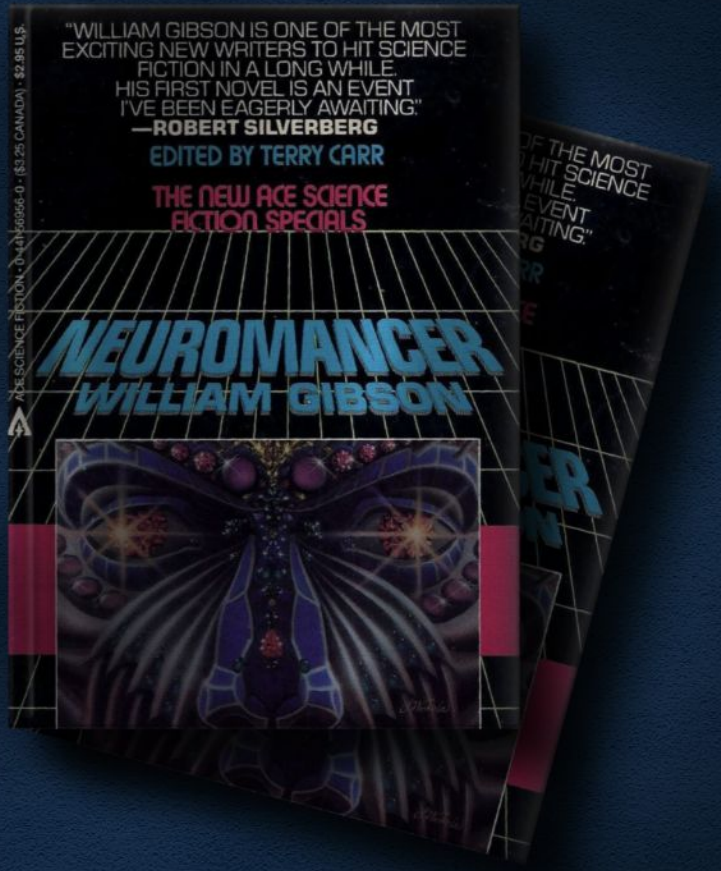


MILEVA
SECURITY LABS

WHAT IS AI?






"WILLIAM GIBSON IS ONE OF THE MOST EXCITING NEW WRITERS TO HIT SCIENCE FICTION IN A LONG WHILE. HIS FIRST NOVEL IS AN EVENT I'VE BEEN EAGERLY AWAITING."
—ROBERT SILVERBERG
EDITED BY TERRY CARR
THE NEW ACE SCIENCE FICTION SPECIALS
NEUROMANCER
WILLIAM GIBSON
ACE SCIENCE FICTION - 0-441-56956-0 - $3.25 CANADA) - $2.95 U.S.

MILEVA SECURITY LABS

AI IN OFFENSIVE SECURITY

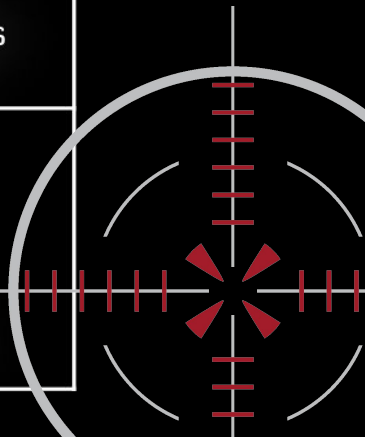
Ethical Hackers (Penetration Testers)	Red Teamers	Exploit Developers
Malware Analysts	Vulnerability Researchers	Social Engineers



Mileva Security Labs

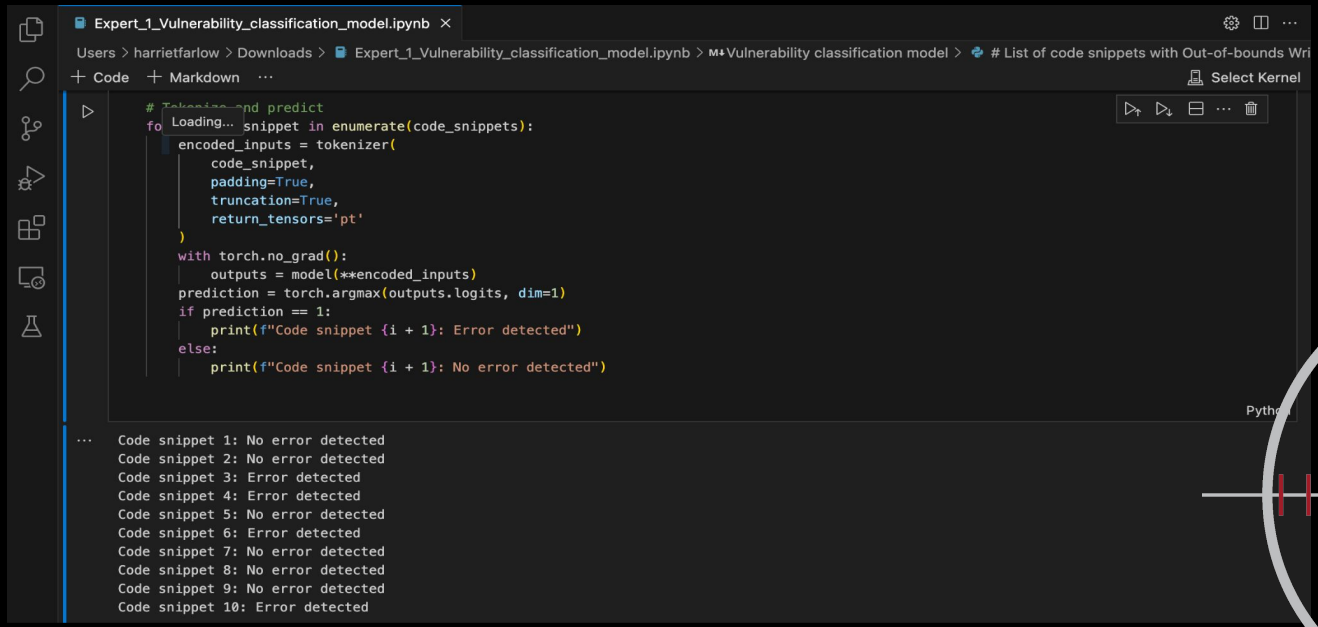
AI IN DEFENSIVE SECURITY

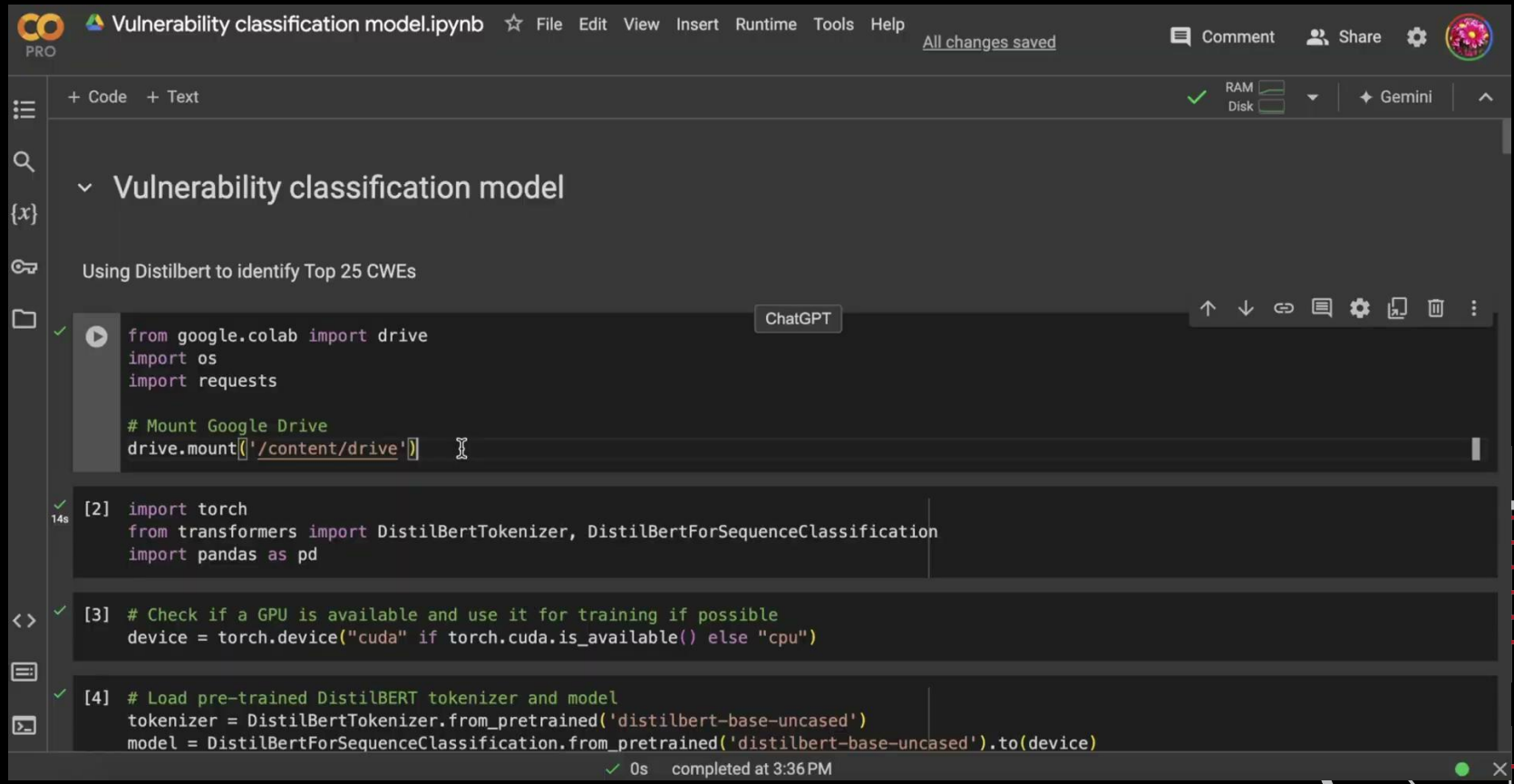
Security Analysts (Defensive)	Network Security Engineers	Security Architects (Defensive)
Incident Responders	Security Operations Center (SOC) Analysts	Threat Hunters



MILEVA SECURITY LABS

DEMO 1: AI FOR CYBER DEFENSE




Vulnerability classification model.ipynb
File Edit View Insert Runtime Tools Help
All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Gemini
Vulnerability classification model
Using Distilbert to identify Top 25 CWEs
ChatGPT
from google.colab import drive
import os
import requests
# Mount Google Drive
drive.mount('/content/drive')
[2] import torch
from transformers import DistilBertTokenizer, DistilBertForSequenceClassification
import pandas as pd
[3] # Check if a GPU is available and use it for training if possible
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
[4] # Load pre-trained DistilBERT tokenizer and model
tokenizer = DistilBertTokenizer.from_pretrained('distilbert-base-uncased')
model = DistilBertForSequenceClassification.from_pretrained('distilbert-base-uncased').to(device)
0s completed at 3:36 PM

MILEVA SECURITY LABS

REFERENT OBJECT



Mileva Security Labs

CASE STUDY: HACKING WITH AI



MILEVA
SECURITY LABS

AI-DRIVEN RECON

3764BG_376N

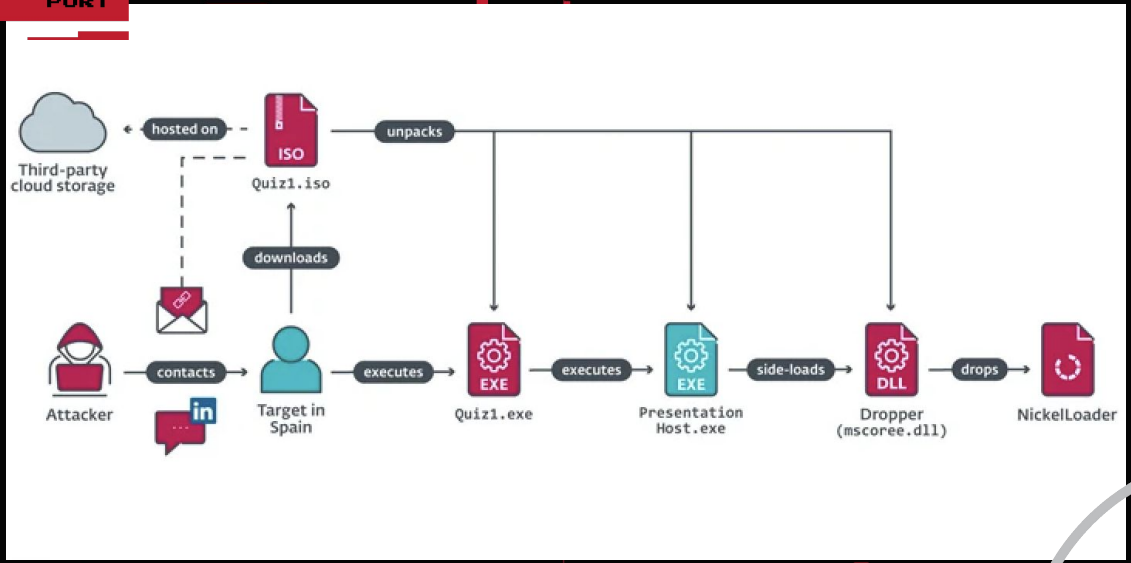
PORT

SYS_478

KILL CHAIN

DONE

VID_37_DGFD




MILEVA
SECURITY LABS
ADVANCED PERSISTENT THREATS
EQUATION GROUP
Origin
United States, 2001
Primary Targets
Iran, Syria and Afganistan
Weapon of Choice
Spyware
MACHETE
Origin
South America, 2010
Primary Targets
Venezuela, Columbia, Nicaragua and Ecuador
Weapon of Choice
Phishing
MYTHIC LEOPARD
Origin
Pakistan, 2016
Primary Targets
India
Weapon of Choice
Social engineering
CHARMING KITTEN
Origin
Iran, 2014
Primary Targets
Israel, Iran, US and UK
Weapon of Choice
Hacking email accounts
LAZARUS GROUP
Origin
North Korea, 2009
Primary Targets
South Korea and US
Weapon of Choice
Ransomware
FANCY BEAR
Origin
Russia, 2004
Primary Targets
US and Germany
Weapon of Choice
Spear-phishing
DYNAMITE PANDA
Origin
China, 2009
Primary Targets
US
Weapon of Choice
Trojan ransomware
OCEANLOTUS
Origin
Vietnam, 2014
Primary Targets
Laos, Philippines, Vietnam and Cambodia
Weapon of Choice
Malware
ELFIN
Origin
Iran, 2013
Primary Targets
Saudi Arabia and US
Weapon of Choice
Shamoon

Mileva
SECURITY LABS

DIRECTION OF PROTECTION



AI for Security

VS



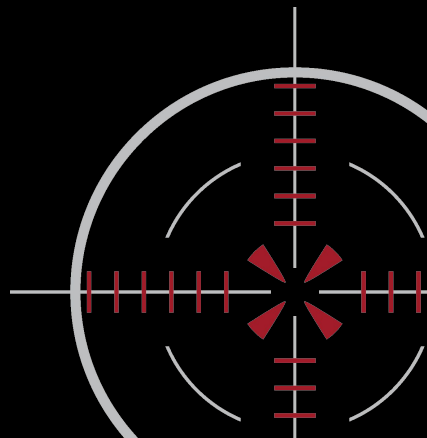
Security of AI



Mileva
SECURITY LABS

HOW TO HACK AN AI





MILEVA
SECURITY LABS

COMPUTER VISION

3764BG_376N

PORT

SYS_478

AUTONOMOUS VEHICLE USE CASE

//////////////////DONE

VID_37_DGFD



MILEVA SECURITY LABS

ADVERSARIAL MACHINE LEARNING



"PANDA"

57.7% confidence

+ .007 x



NOISE

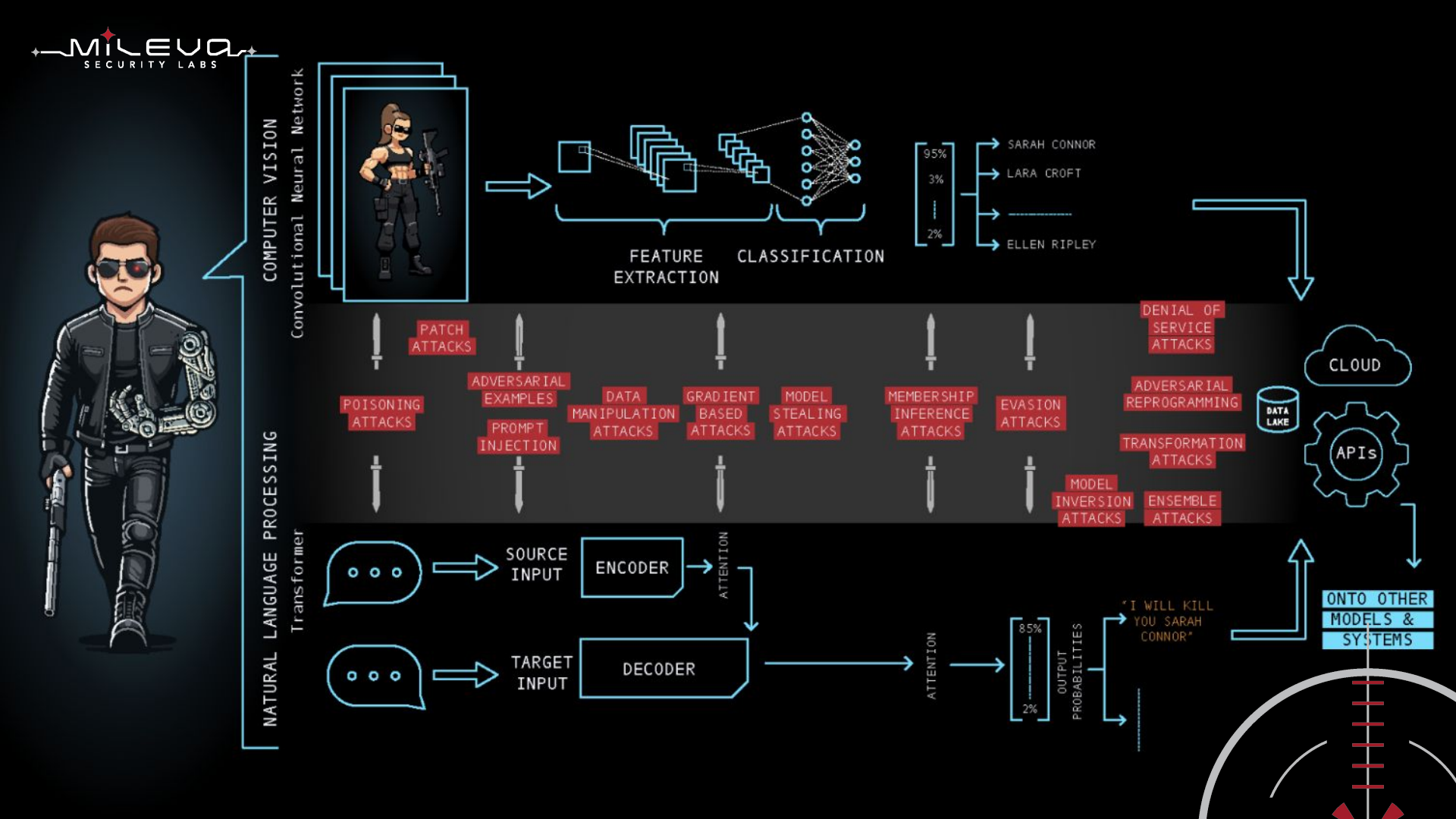
8.2% confidence

=



"GIBBON"

99.3% confidence


MILEVA
SECURITY LABS
COMPUTER VISION
Convolutional Neural Network
FEATURE EXTRACTION
CLASSIFICATION
95%
3%
2%
SARAH CONNOR
LARA CROFT
ELLEN RIPLEY
PATCH ATTACKS
POISONING ATTACKS
ADVERSARIAL EXAMPLES
PROMPT INJECTION
DATA MANIPULATION ATTACKS
GRADIENT BASED ATTACKS
MODEL STEALING ATTACKS
MEMBERSHIP INFERENCE ATTACKS
EVASION ATTACKS
DENIAL OF SERVICE ATTACKS
ADVERSARIAL REPROGRAMMING
TRANSFORMATION ATTACKS
MODEL INVERSION ATTACKS
ENSEMBLE ATTACKS
DATA LAKE
CLOUD
APIs
NATURAL LANGUAGE PROCESSING
Transformer
SOURCE INPUT
ENCODER
ATTENTION
TARGET INPUT
DECODER
ATTENTION
85%
2%
OUTPUT PROBABILITIES
"I WILL KILL YOU SARAH CONNOR"
ONTO OTHER MODELS & SYSTEMS

MILEVA
SECURITY LABS

MISSION IMPOSSIBLE

SYS_478

HACKING
SURVEILLANCE
AI

//////////////////DONE

VID_37_DGFD



MILEVA
SECURITY LABS

HACKING FACIAL RECOGNITION

3764BG_376N

PORT

SYS_478

NO MATCH FOUND

//////////////////DONE

VID_37_DGFD



Demo - facial recognition.ipynb ☆ File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text RAM Disk Gemini

Creating DAR jewellery

```
!pip install tensorflow-addons
!pip install tensorflow
```

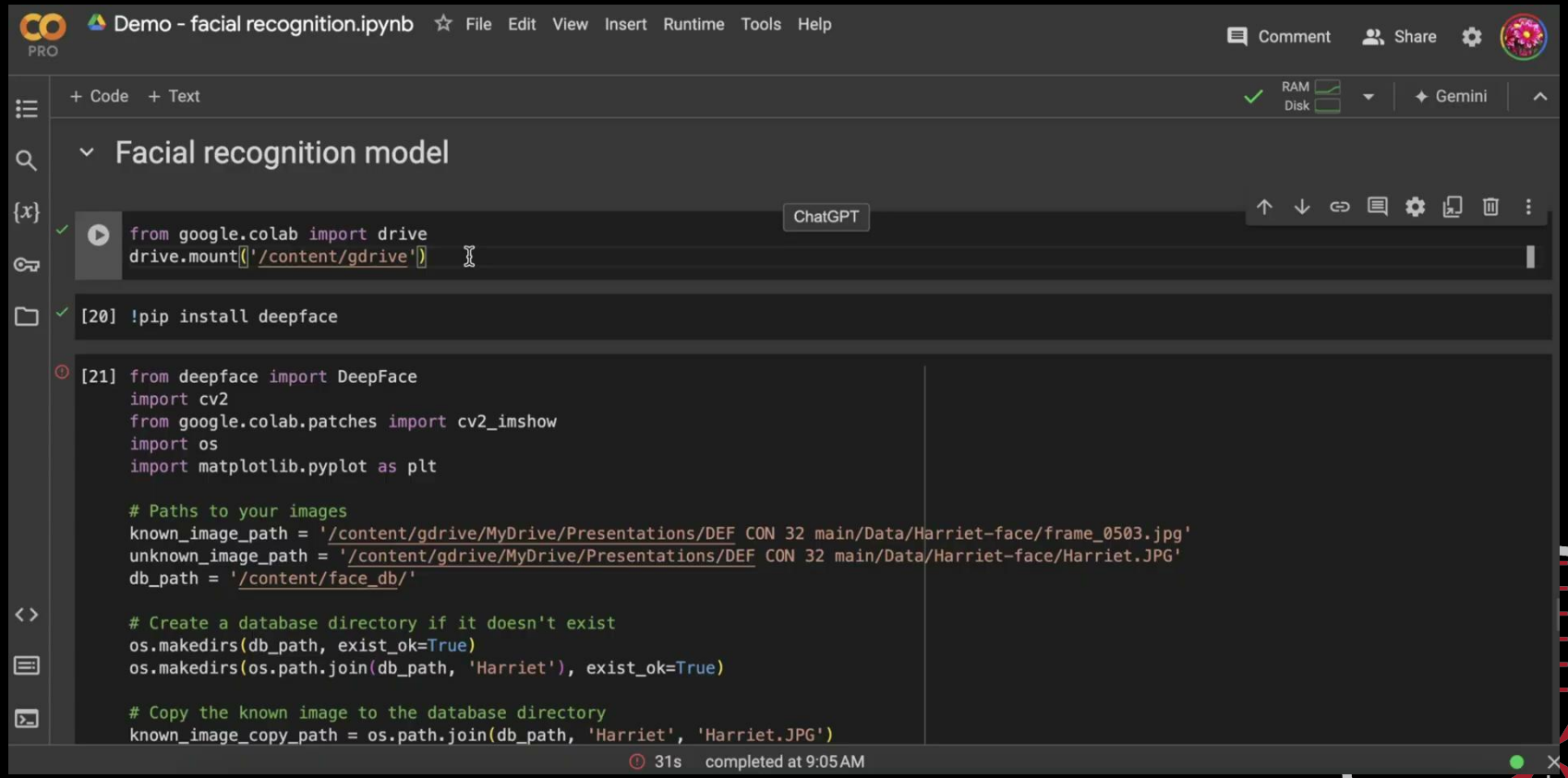
```
[42] from google.colab import drive
     drive.mount('/content/gdrive')

     import numpy as np
     import tensorflow as tf
     from tensorflow.keras.preprocessing import image
     from tensorflow.keras.applications.inception_resnet_v2 import InceptionResNetV2, preprocess_input
     import matplotlib.pyplot as plt
     from skimage.draw import disk
     from scipy.ndimage import zoom
```

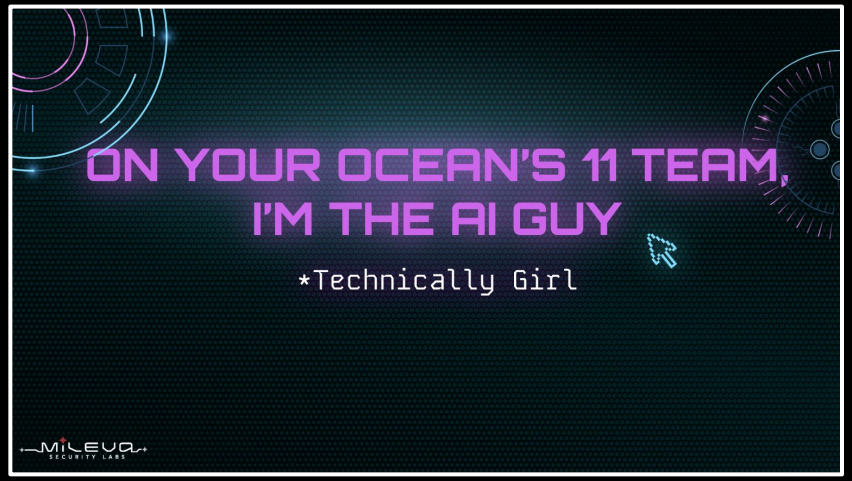
```
[43] # Load the InceptionResNetV2 model pre-trained on ImageNet
     base_model = InceptionResNetV2(include_top=False, pooling='avg', input_shape=(160, 160, 3))

     # Add a dense layer to obtain embeddings
     inputs = tf.keras.Input(shape=(160, 160, 3))
     x = base_model(inputs, training=False)
     outputs = tf.keras.layers.Dense(128, activation=None)(x)
     facenet_model = tf.keras.Model(inputs, outputs)
```

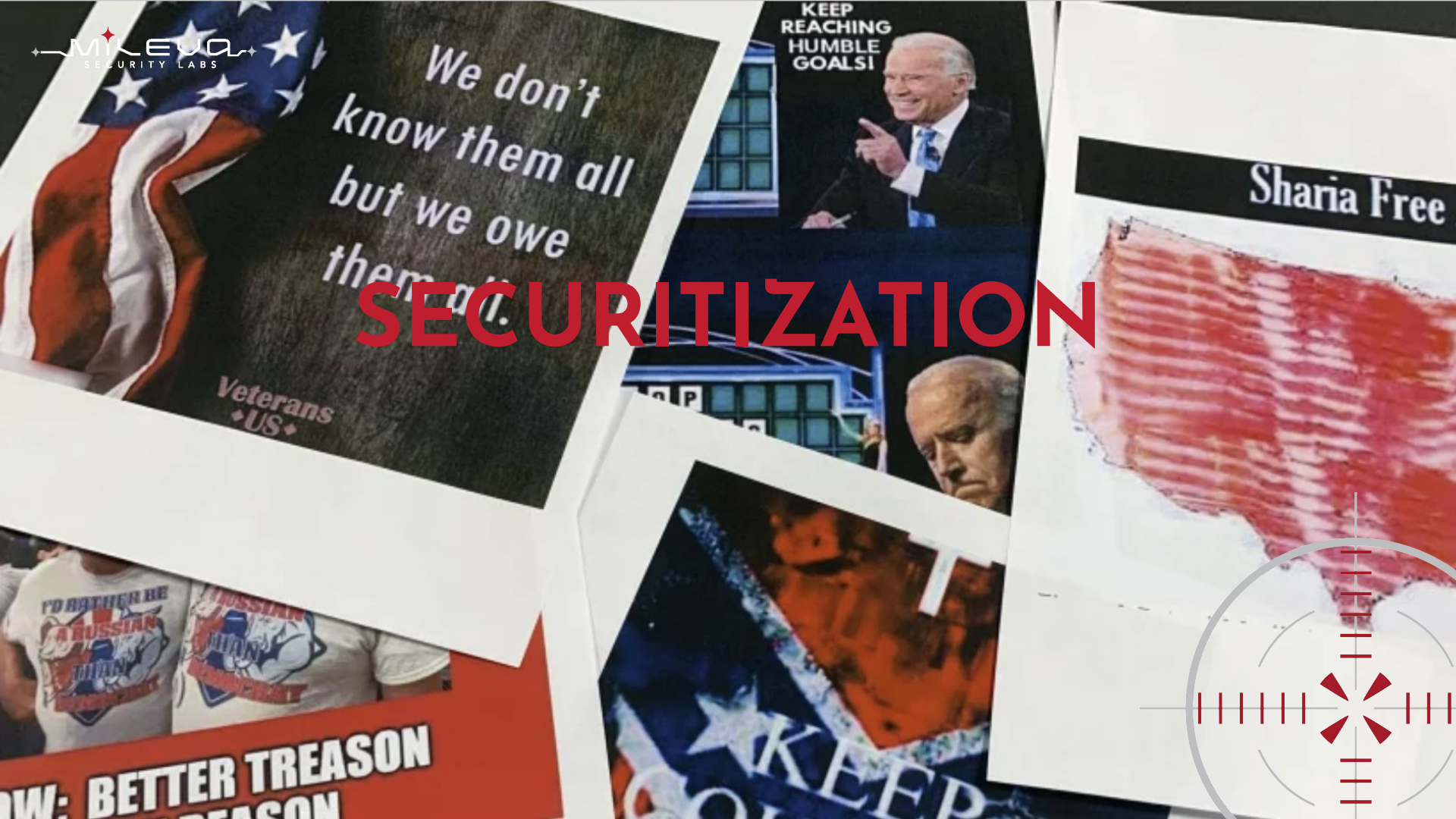
✓ 1s completed at 9:17 AM


Demo - facial recognition.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code + Text
RAM
Disk
Gemini
Facial recognition model
ChatGPT
from google.colab import drive
drive.mount('/content/gdrive')
[20] !pip install deepface
[21] from deepface import DeepFace
import cv2
from google.colab.patches import cv2_imshow
import os
import matplotlib.pyplot as plt
# Paths to your images
known_image_path = '/content/gdrive/MyDrive/Presentations/DEF CON 32 main/Data/Harriet-face/frame_0503.jpg'
unknown_image_path = '/content/gdrive/MyDrive/Presentations/DEF CON 32 main/Data/Harriet-face/Harriet.JPG'
db_path = '/content/face_db/'
# Create a database directory if it doesn't exist
os.makedirs(db_path, exist_ok=True)
os.makedirs(os.path.join(db_path, 'Harriet'), exist_ok=True)
# Copy the known image to the database directory
known_image_copy_path = os.path.join(db_path, 'Harriet', 'Harriet.JPG')
31s completed at 9:05AM

SEE MY OTHER DEF CON TALK







SECURITIZATION

MILEVA SECURITY LABS

POLICY



MILEVA
SECURITY LABS

PEOPLE



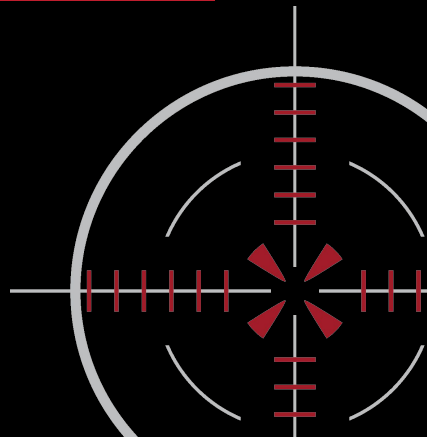
MILEVA SECURITY LABS

DISCUSSION

IS AI BEING SECURITIZED

WHAT ADVERSARIAL USES OF AI ARE YOU SEEING

WHAT DO YOU WISH MORE PEOPLE OUTSIDE YOUR FIELD KNEW



MILEVA SECURITY LABS

STAY IN TOUCH

@HarrietHacks

