ON YOUR OCEAN'S 11 TEAM, I'M THE AI GUY

*Technically Girl

Harriet Farlow

MILEVA
SECURITY LABS

MILEVA SECURITY LABS

OUR OBJECTIVES

Hack the Casino

- Casino OSINT and recon
- Hack facial recognition AI
- Reflect on AI Security



Mileva Security Labs

WHO AM I





Mileva Security Labs

@HarrietHacks

MILEVA SECURITY LABS

WHY THE CASINO?



| All AI can be hacked, but people only care when you lose a lot of money.. publically



MILEVA SECURITY LABS

SOME DISCLAIMERS

- We had Casino Canberra's permission
- These attacks are real but we don't show you how to implement them
- This talk is aimed at cyber folk (don't come at me AI people!)



MILEVA
SECURITY LABS

A BIT ABOUT CASINOS

For the non-gamblers among us

- They make a lot of money, but have a controversial history



MILEVA SECURITY LABS

A BIT ABOUT CASINOS

A BIT OF HEAT

Politics

Money launderers are using online gambling sites to disguise shady funds

Fintrac analyzed suspicious transaction reports related to online gambling between 2016 and 2023

Jim Bronskill · The Canadian Press · Posted: Jan 18, 2024 9:52 AM EST | Last Updated: January 18



Australia news

This article is more than **1 year old**

Crown to pay $450m penalty after failing to comply with anti-money laundering laws

Crown casino and Austrac agree to penalty, which still must be approved by court

- Follow our Australia news live blog for the latest updates
- Get our morning and afternoon news emails, free app or daily news podcast

Australian Associated Press

Tue 30 May 2023 06.05 BST

Share



MILEVA SECURITY LABS

A BIT ABOUT CASINOS

For the non-gamblers among us

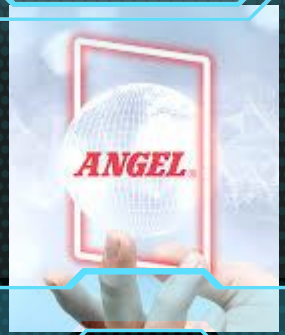
- They make a lot of money, but have been getting a lot of heat
- They use AI



MILEVA
SECURITY LABS

A BIT ABOUT CASINOS







MILEVA SECURITY LABS

A BIT ABOUT CASINOS

For the non-gamblers among us

- They make a lot of money, but have been getting a lot of heat
- They use AI
- Card counting is not illegal, but is considered advantage play

MILEVA SECURITY LABS

HOW TO HACK AN AI



Mileva
SECURITY LABS

ADVERSARIAL MACHINE LEARNING



"PANDA"

57.7% confidence

+ .007 x



NOISE

8.2% confidence

=



"GIBBON"

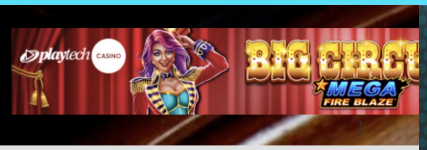
99.3% confidence

MILEVA SECURITY LABS

ALGORITHM HACKING

RANDOM NUMBER GENERATORS

GGB
Global Gaming Business Magazine



***GGB** is committed to providing updated news and analysis on our weekly news site, GGBNews.com.*

Features

SECURITY & SURVEILLANCE: Running the RNG Risk

By Willy Allison *Fri, Jan 20, 2017*

SHARE ON: Twitter Facebook LinkedIn

How computers are taking the luck out of gambling



In the late '80s, a real game-changer took place in casinos that would forever change the landscape and opportunities in the gambling industry. Slot machines went from being mechanical devices to being electronic devices.

In order to provide randomness and bigger payouts, a new computerized process was introduced that utilized a random number generator (RNG). The RNG is an algorithm

MILEVA SECURITY LABS

ALGORITHM HACKING

CARD COUNTING

THE RUNNING COUNT



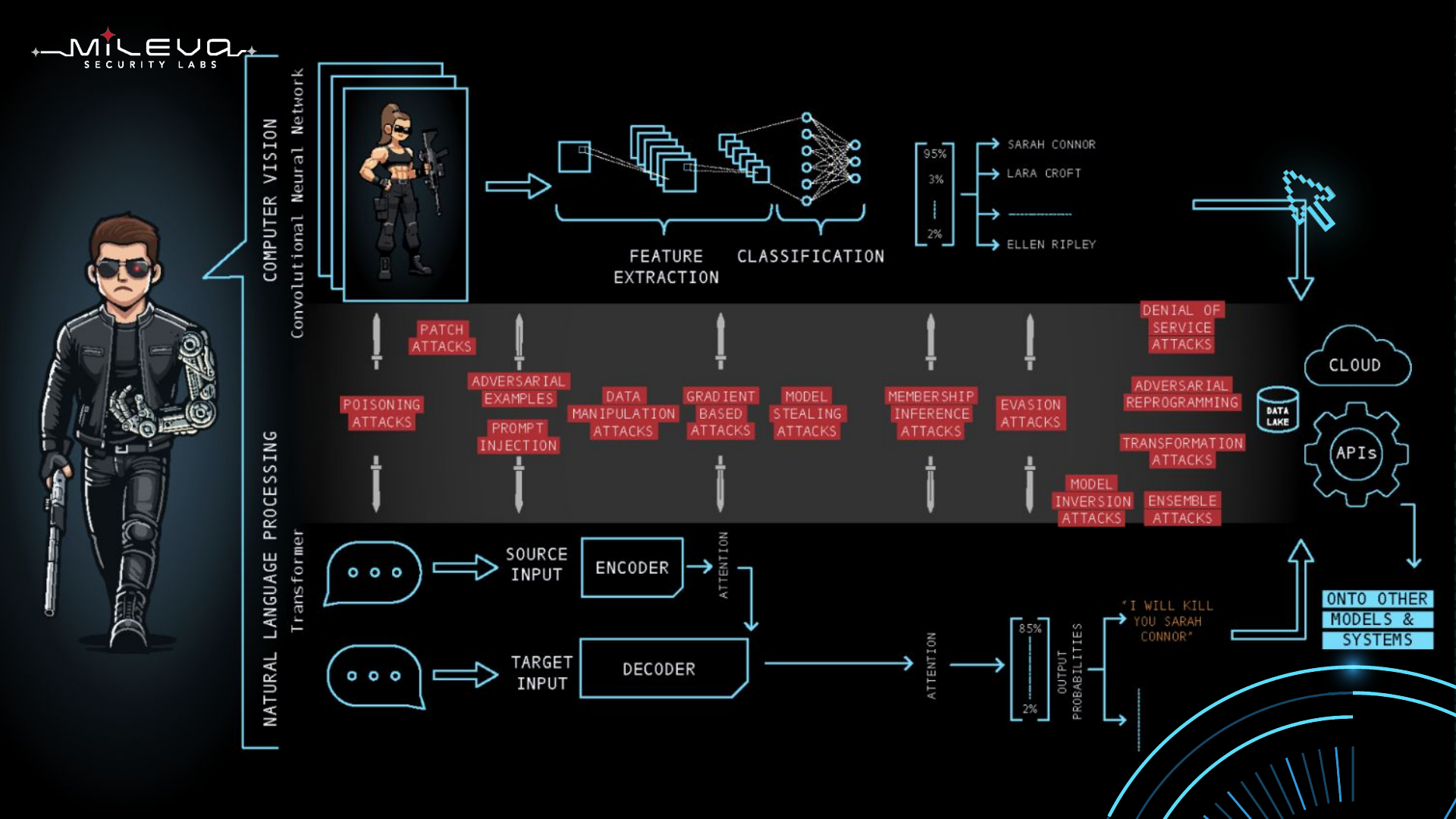
high cards → Value -1



low cards → Value +1



Value 0


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

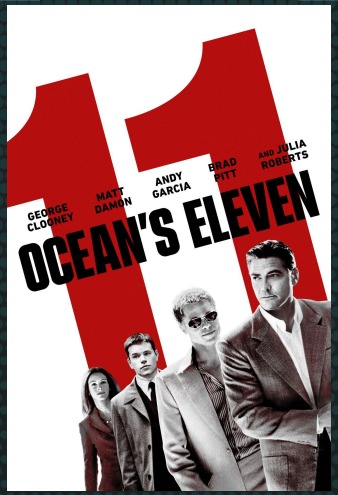
HAS EVERYONE SEEN OCEAN'S 11?



MILEVA SECURITY LABS

HEIST 101







MILEVA SECURITY LABS

HACKING THE CASINO

INTERVIEW CASINO STAFF

CREATE DARs (DISTRIBUTED ADVERSARIAL REGIONS)



VALIDATE AND LESSONS LEARNT

PICK SOME TARGET (SURROGATE) MODELS

DISGUISE DARs AS JEWELLERY

MILEVA SECURITY LABS

STEP 1: RECON

| Tell people I'm a PhD student doing a paper on AI security, and looking to interview experts



STEP 1: RECON



MILEVA SECURITY LABS

INTERVIEW ANALYSIS

50 interviews

43 organizations

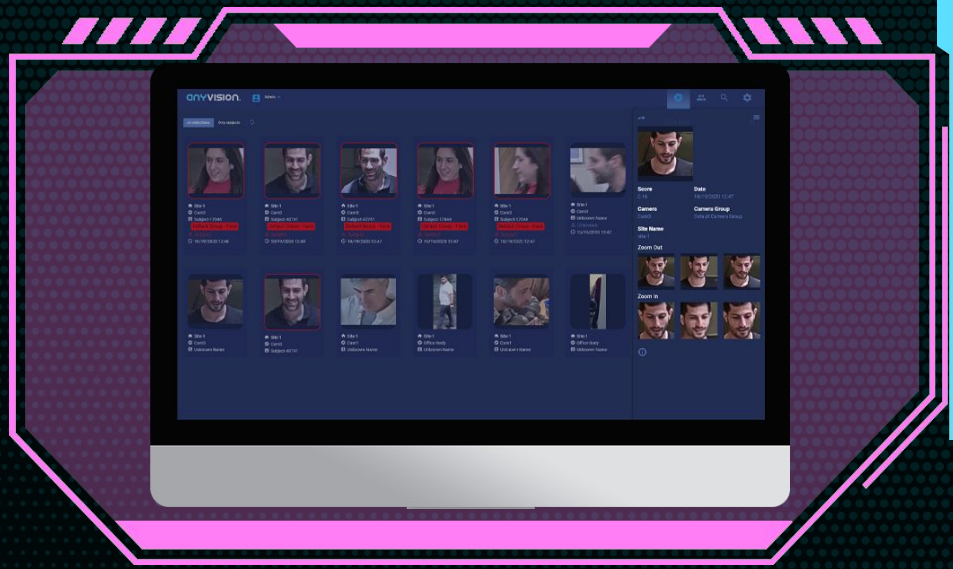
94% could articulate how they **use** AI

8% could articulate how they **secure** their AI

MILEVA SECURITY LABS

CASINOS

FACIAL RECOGNITION






MILEVA
SECURITY LABS
PEOPLE

MILEVA SECURITY LABS

STEP 2: CHOOSE A VICTIM

VGGFace

DeepFace

DeepID

SFace

OpenFace

FaceNet-123d

FaceNet-512d

ArcFace

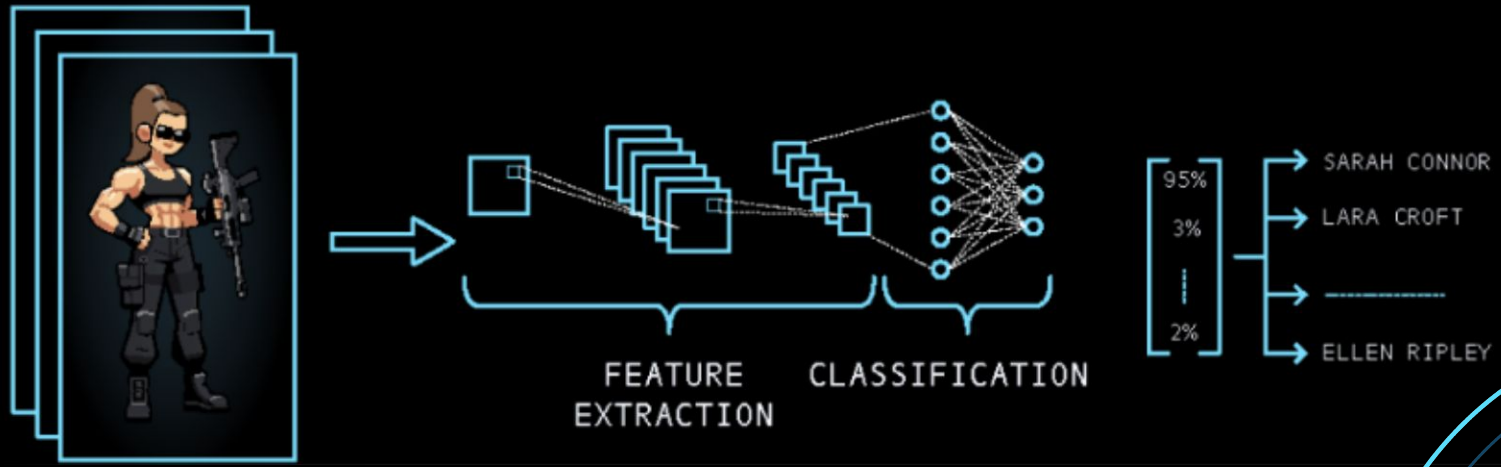
GhostFaceNet

Custom (from scratch)



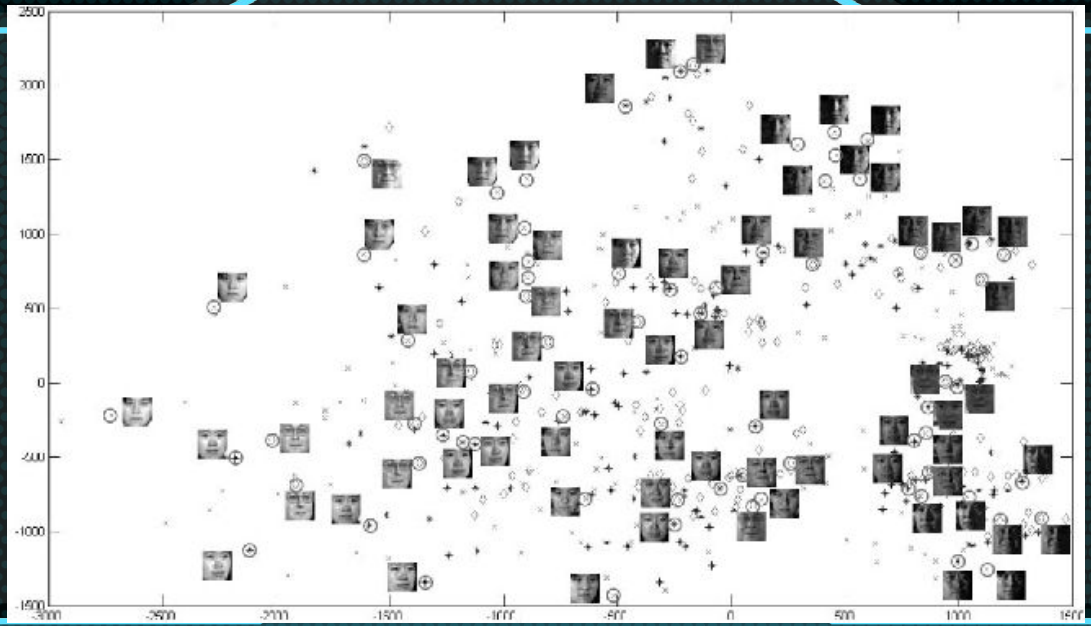
MILEVA SECURITY LABS

CONVOLUTIONAL NEURAL NETWORK



MILEVA SECURITY LABS

EMBEDDING SPACE



MILEVA
SECURITY LABS

MODEL EVALS

DeepFace	FaceNet	VGG-Face	ArcFace	DeepID2+
97.35%	99.63%	98.95%	99.83%	99.47%

MILEVA
SECURITY LABS

95%



MILEVA
SECURITY LABS

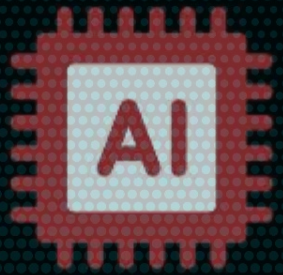
CONVERGENCE



MILEVA
SECURITY LABS



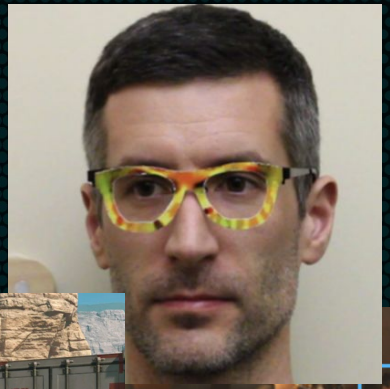




same data + same training = same model

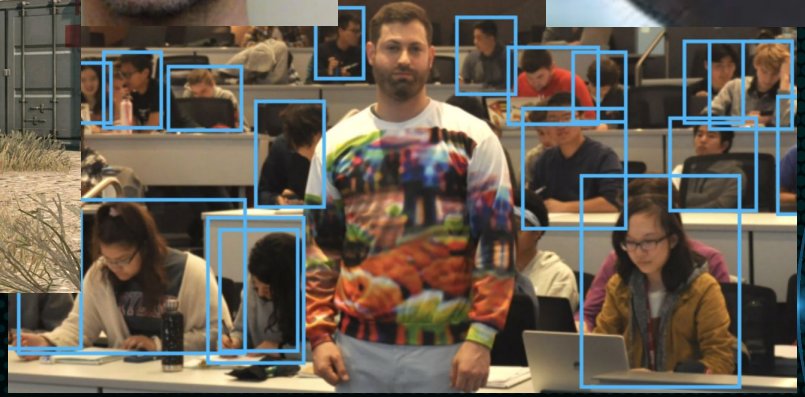
MILEVA SECURITY LABS

STEP 3: CREATE DARS



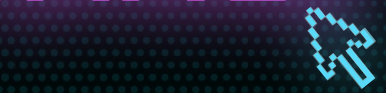






MILEVA SECURITY LABS

STEP 3: CREATE DARS



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

1

Identify most influential regions by convolving an n diameter pixel region



Influence measured based on the maximum decrease in confidence based on target model's predictions

2

Apply DARs at regions identified in step 1 and modify DAR parameters

Universal Adversarial Perturbation (UAP)



Projected Gradient Descent (PGD)



Size (n)	2, 6, 8, 10, 14, 18
Number	1, 2, 3, 4

3

Assess the impact of DARs against different benchmarks

	White box	Black box
Models	MobileNet -V2	NasNet-Mobile DenseNet 169
Data	ImageNet	COIL

4 test different shapes, filters and contexts

STEP 3: CREATE DARS



87% YACHT

MILEVA SECURITY LABS

STEP 3: CREATE DARS



63% BIRD

MILEVA SECURITY LABS

STEP 3: CREATE DARS



77% CAR

MILEVA SECURITY LABS

STEP 3: CREATE DARS



87% YACHT

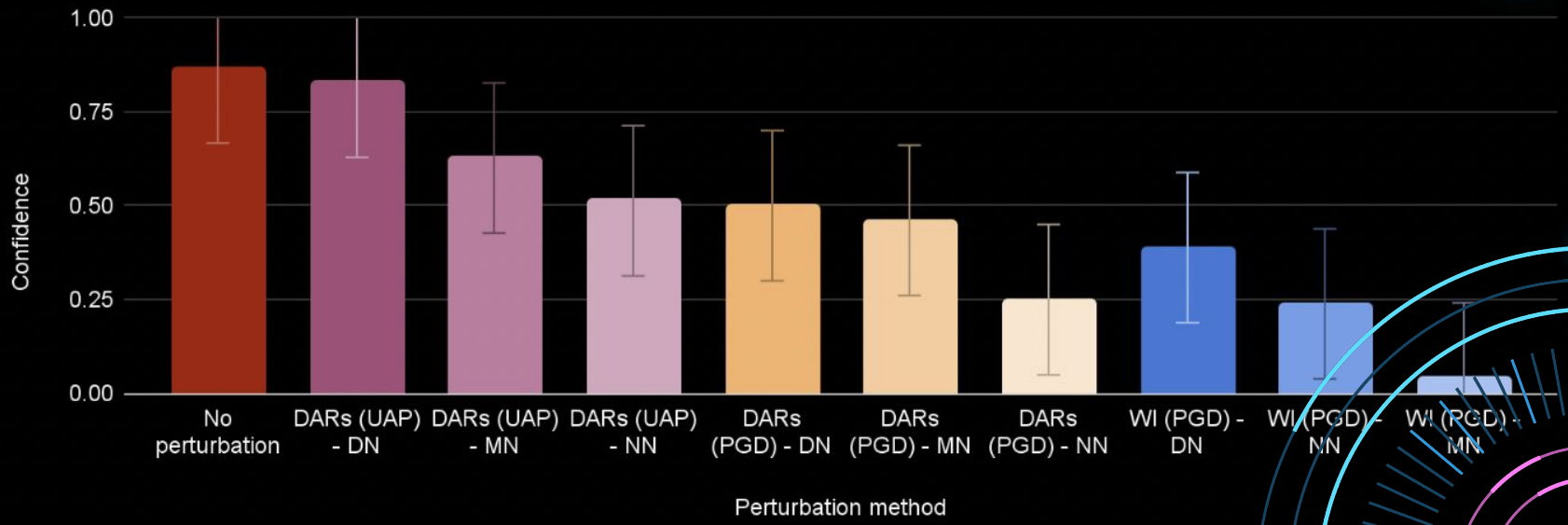


63% BIRD



77% CAR

STEP 3: CREATE DARS



Mileva
SECURITY LABS

STEP 3: CREATE DARS



MILEVA SECURITY LABS

STEP 3: CREATE DARS

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
```

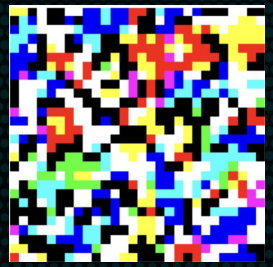
```
[43] # Load the InceptionResNetV2 model pre-trained on ImageNet
     base_model = InceptionResNetV2(include_top=False, pooling='avg', input_shape=(160, 160, 3))

     # 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

MILEVA SECURITY LABS

STEP 4: DISGUISE AS JEWELLERY



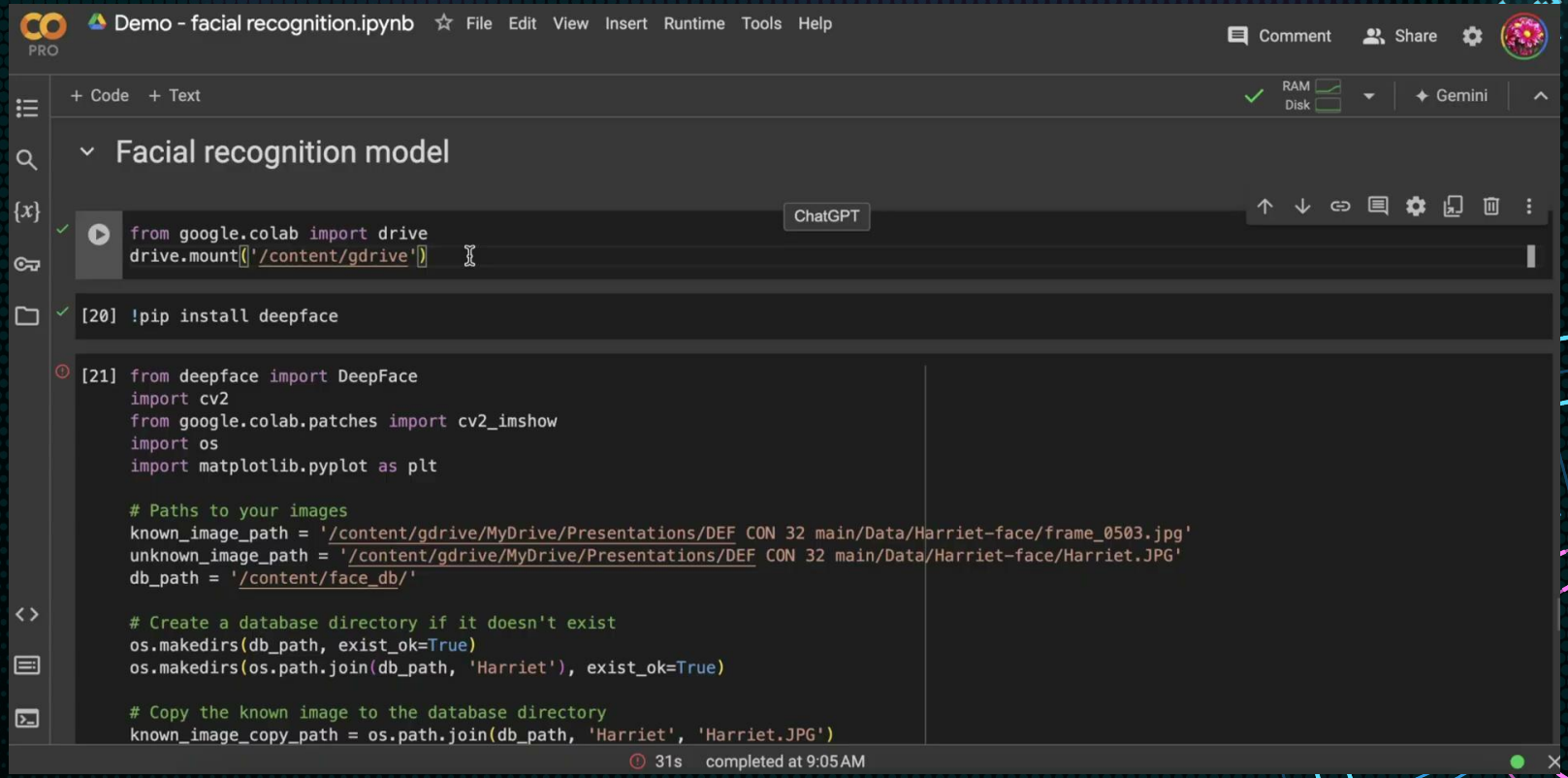






MILEVA SECURITY LABS

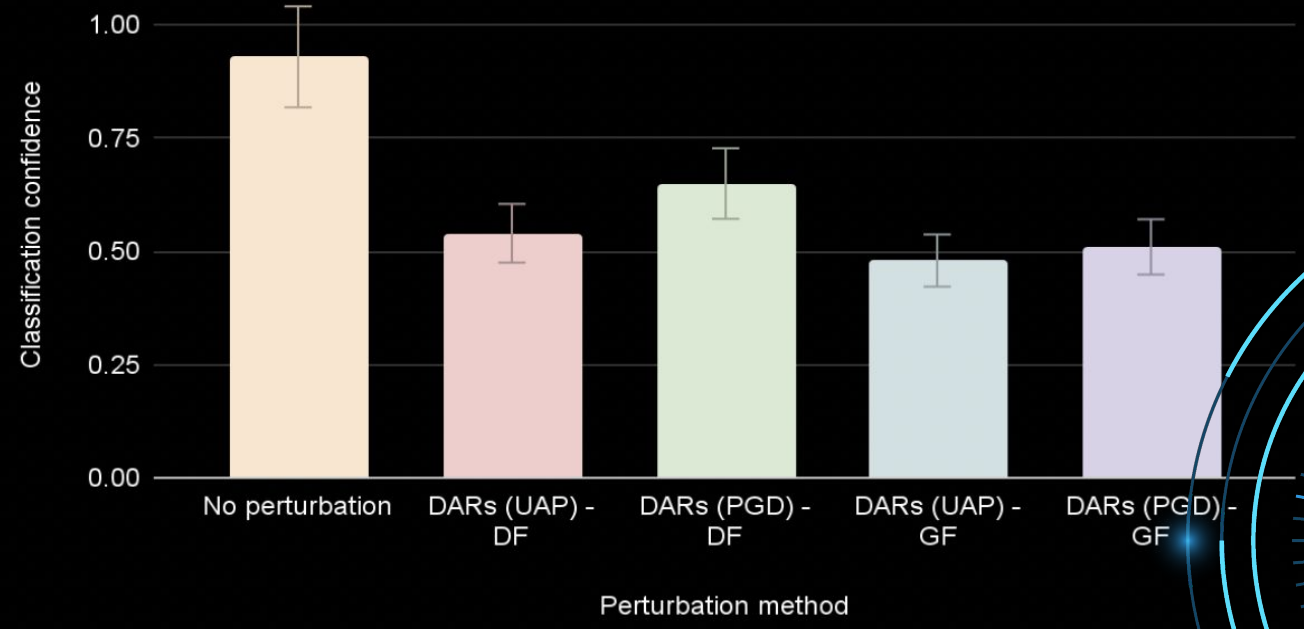
STEP 4: DISGUISE



Winner winner, chicken dinner



STEP 5: TEST


1.00
0.75
0.50
0.25
0.00
Classification confidence
No perturbation
DARs (UAP) - DF
DARs (PGD) - DF
DARs (UAP) - GF
DARs (PGD) - GF
Perturbation method

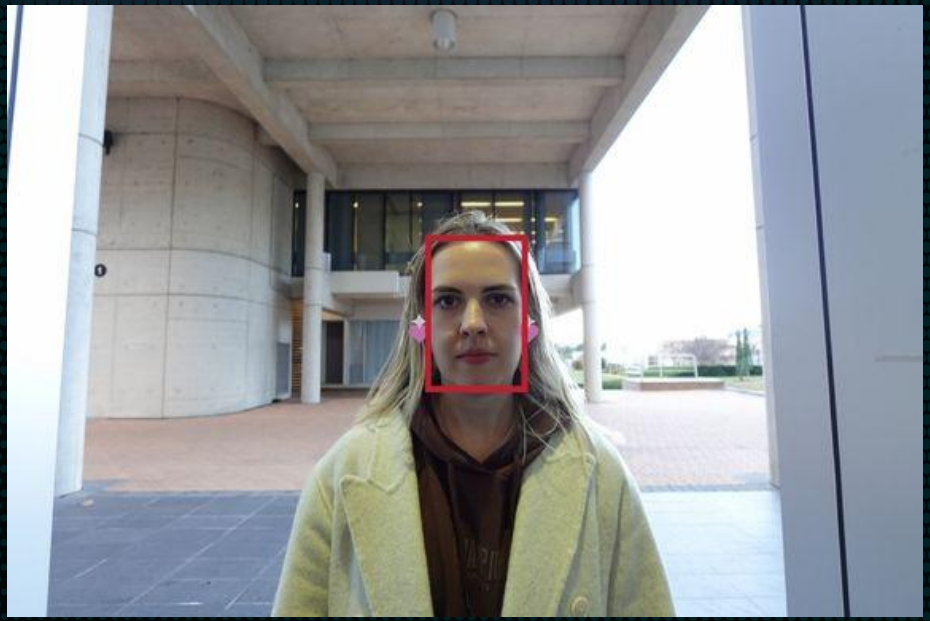
MILEVA
SECURITY LABS

40.4%



MILEVA SECURITY LABS

STEP 4: DISGUISE AS JEWELLERY



MILEVA SECURITY LABS

CONTEXT



Mileva Security Labs

STEP 5: TEST



Mileva
SECURITY LABS

STEP 6: BOOK A TRIP









MILEVA
SECURITY LABS

WHAT AM I TRYING TO STEAL?

(What is the target?)



MILEVA
SECURITY LABS

WHAT DOES IT MEAN TO 'HACK' AN AI?

(When the system is inherently probabilistic?)



MILEVA
SECURITY LABS

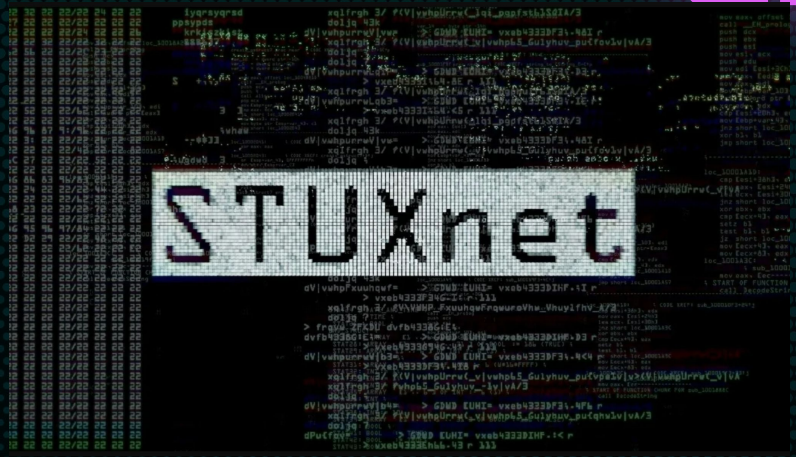
WHAT ARE WE TRYING TO PROTECT?

(What risk management practices are in place?)



MILEVA
SECURITY LABS

PART OF THE KILL CHAIN





MILEVA
SECURITY LABS

40.4%



MILEVA
SECURITY LABS

100%

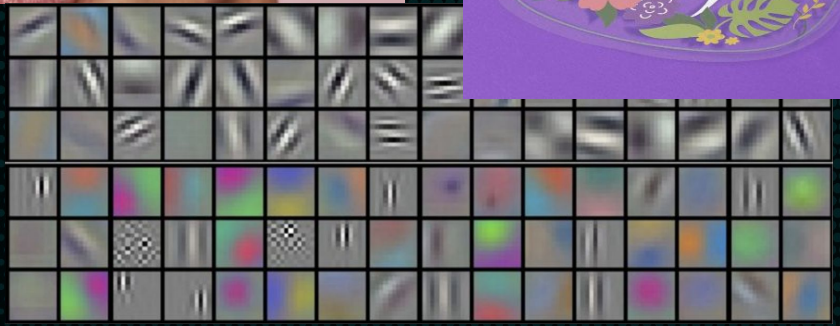


MILEVA
SECURITY LABS

PART OF THE KILL CHAIN







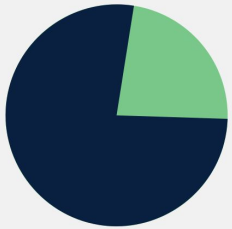
MILEVA SECURITY LABS

FACIAL RECOGNITION IN NON-HEIST SCENARIOS



MILEVA SECURITY LABS

AI SECURITY IS REAL



77%

of companies reported identifying breaches to their AI in the past year. The remaining were uncertain whether their AI models had seen an attack.

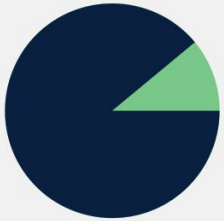
30% of IT leaders have deployed a manual defense for adversarial attacks on their existing AI, while just

14% are planning and testing for such attacks.

30% Only 30% have deployed technology for model theftjacking, with

20% planning and testing for this threat.

HIDDENLAYER



96%

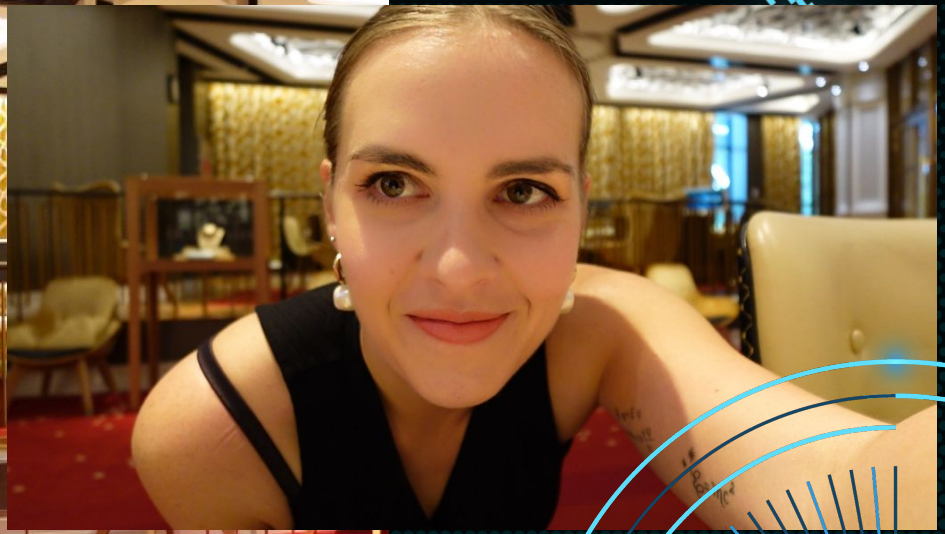
of IT leaders expressed that their AI projects are critical or important to revenue generation over the next 18 months.

~~STEP 7: EVADE FACIAL RECOGNITION AI~~

STEP 7: PEOPLE AND PROCESS

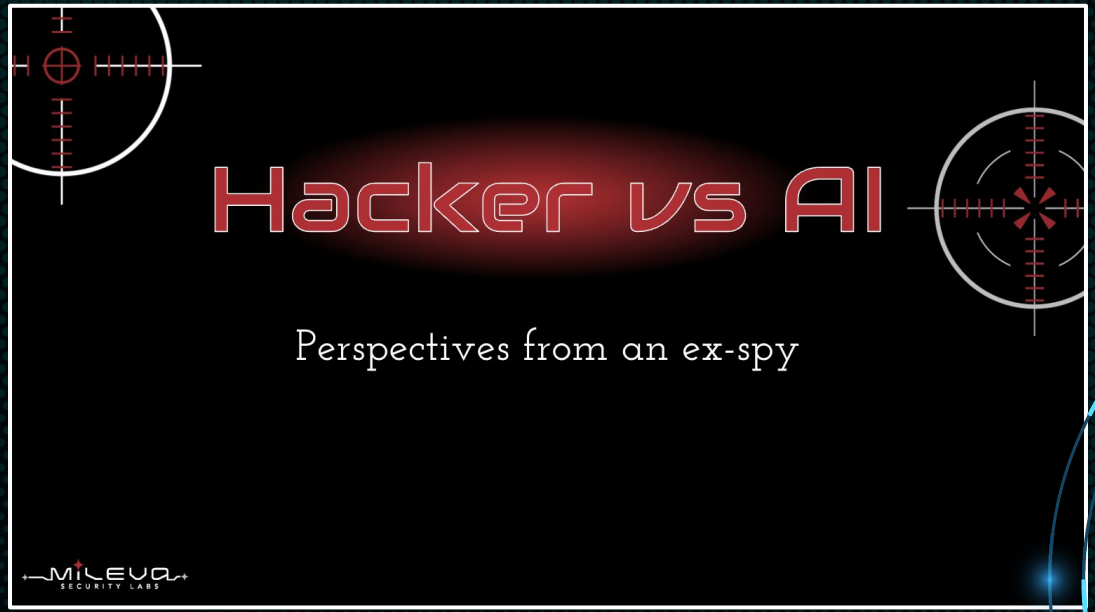
MILEVA
SECURITY LABS





MILEVA SECURITY LABS

SEE MY OTHER TALK



MILEVA SECURITY LABS

STAY IN TOUCH



@HarrietHacks

I'm also looking for survey participants..