# Data poisoning won’t save you from facial recognition

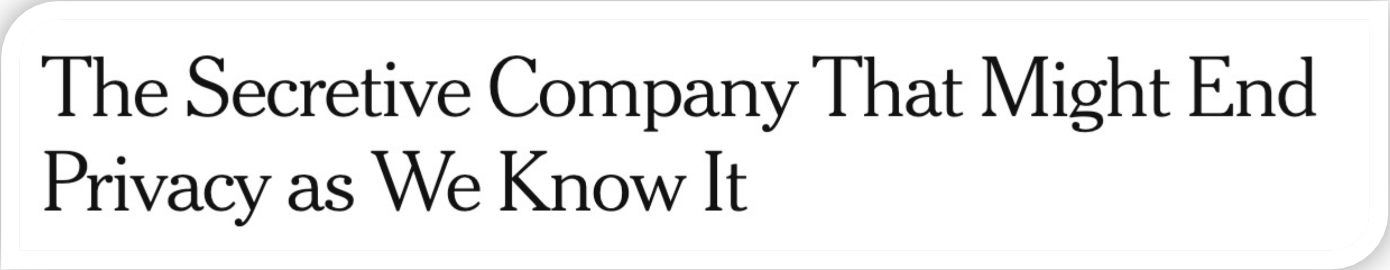
**Florian Tramèr**
Stanford University

CPVR 2021 Workshop on Media Forensics

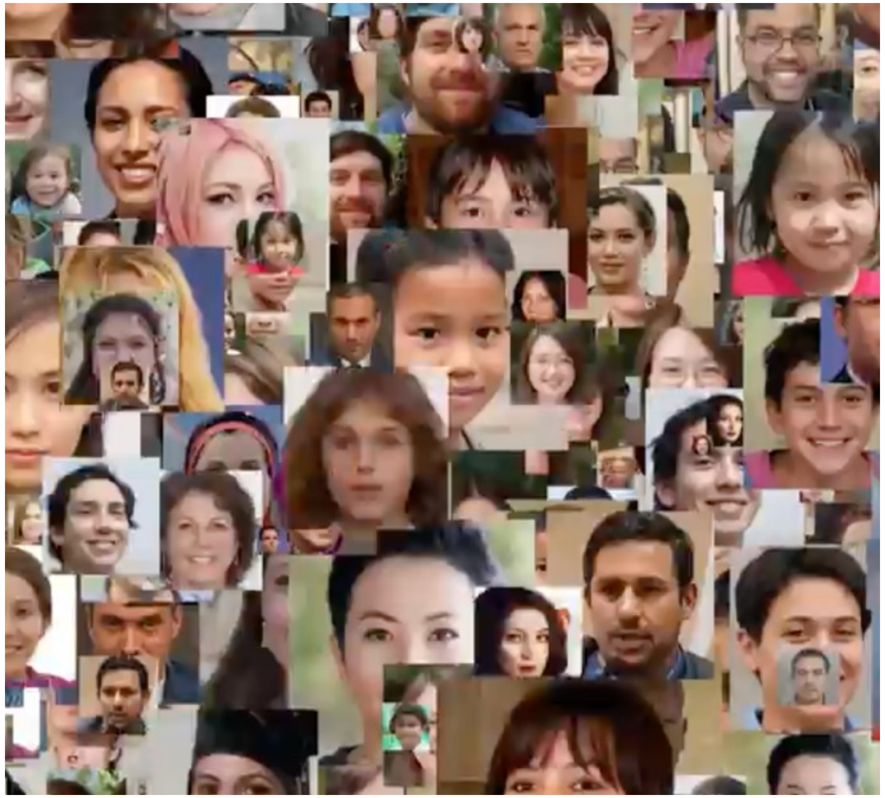
(joint work with Evani Radiya-Dixit)

# You can't hide from Big Brother.

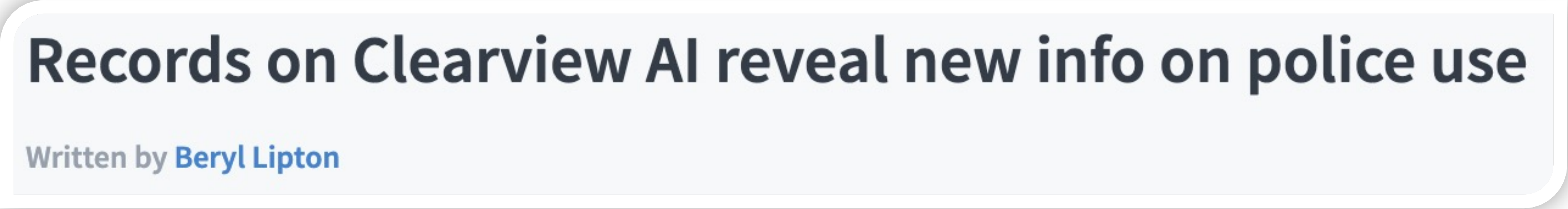
Clearview.ai



The Secretive Company That Might End Privacy as We Know It

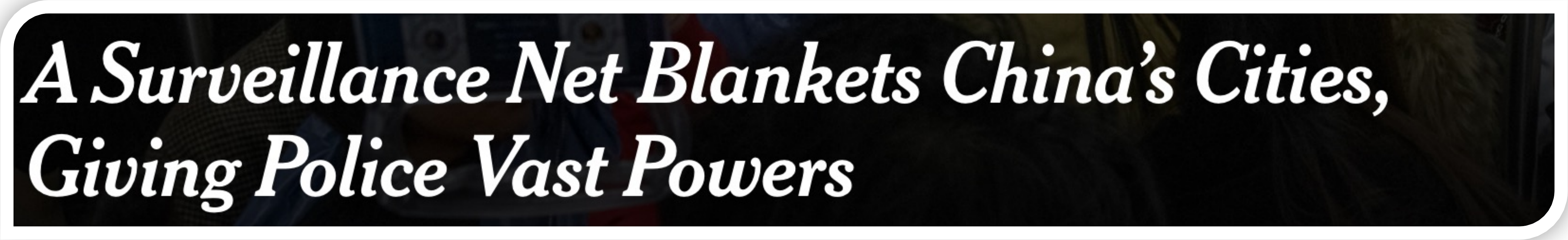


# You can't hide from Big Brother.



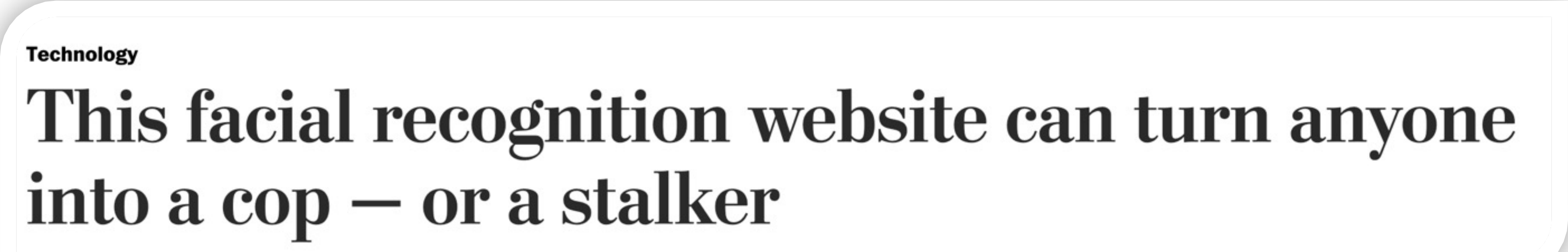
Records on Clearview AI reveal new info on police use

Written by Beryl Lipton



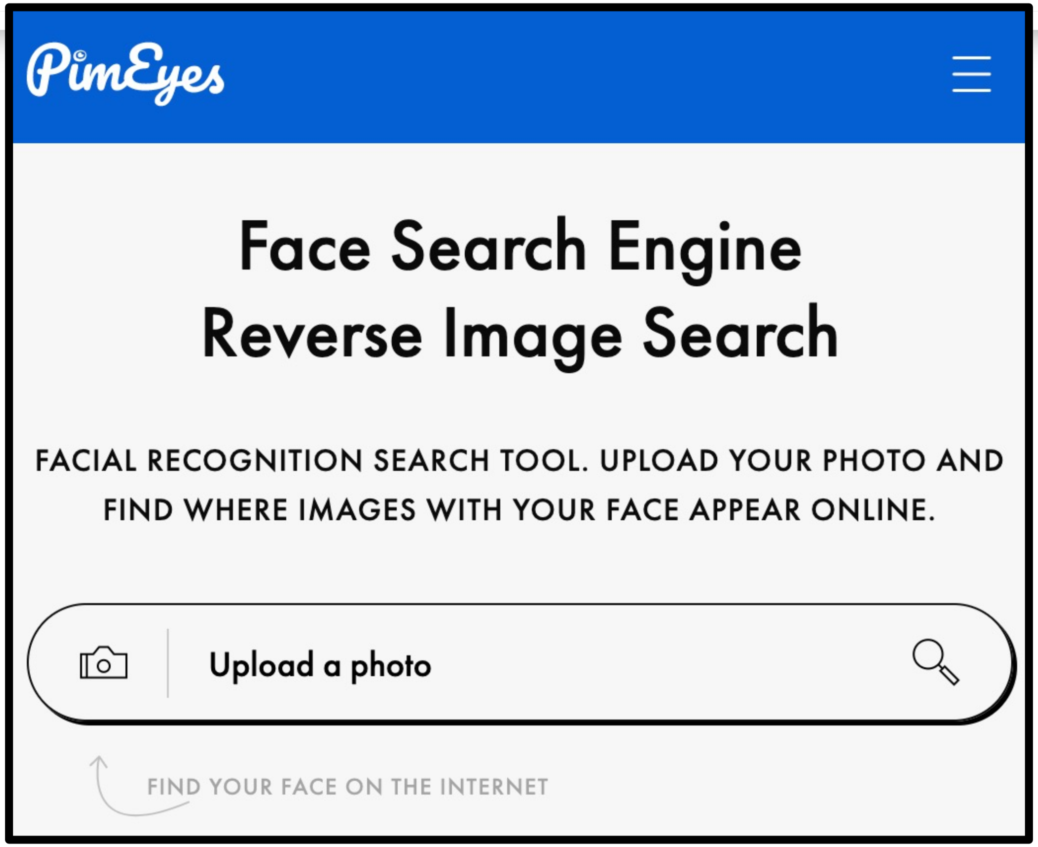
A Surveillance Net Blankets China's Cities, Giving Police Vast Powers

# You can't hide from anyone.



Technology

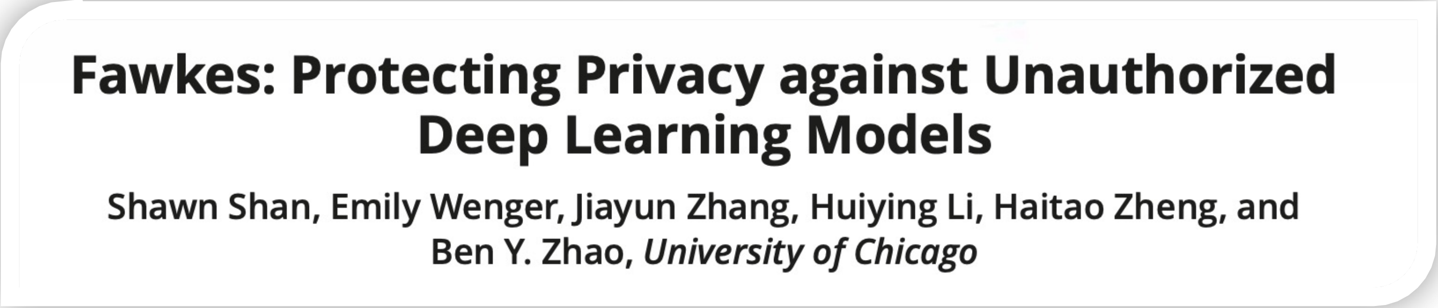
This facial recognition website can turn anyone into a cop — or a stalker



Face Search Engine
Reverse Image Search

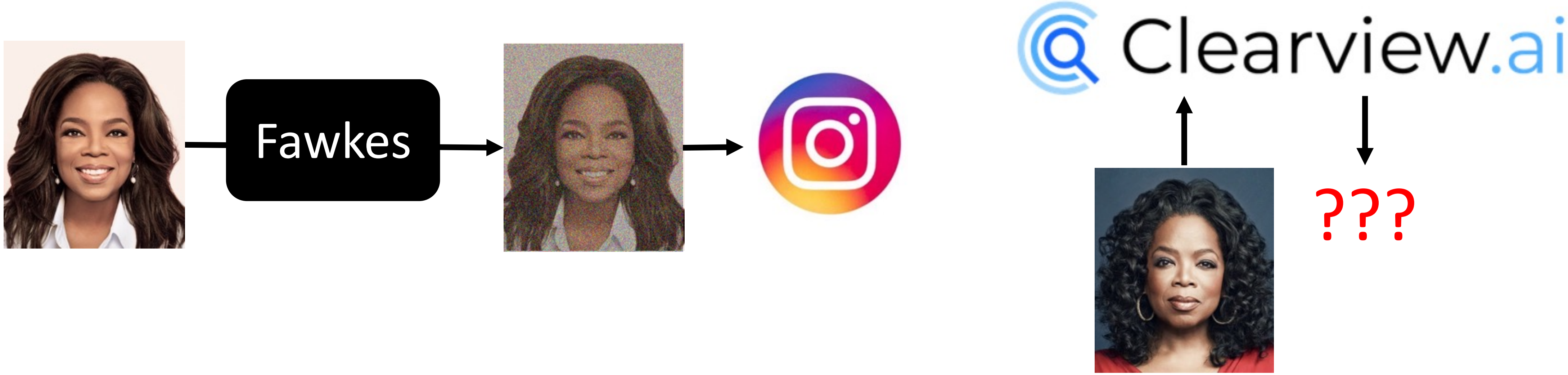
FACIAL RECOGNITION SEARCH TOOL. UPLOAD YOUR PHOTO AND FIND WHERE IMAGES WITH YOUR FACE APPEAR ONLINE.

# Image-perturbation tools promise to defeat facial recognition.

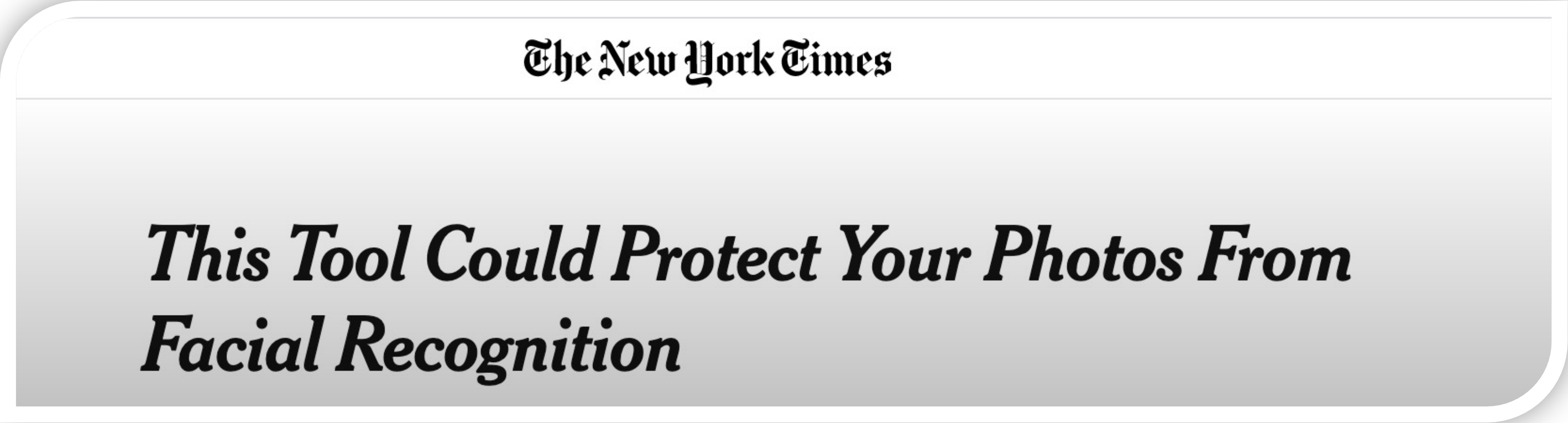


**Fawkes: Protecting Privacy against Unauthorized Deep Learning Models**

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao, *University of Chicago*

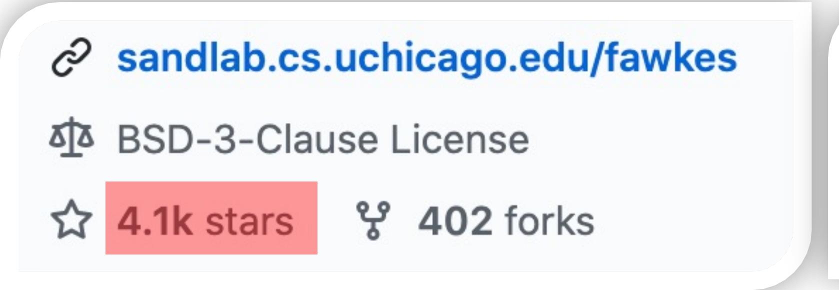


# Image-perturbation tools promise to defeat facial recognition.



The New York Times

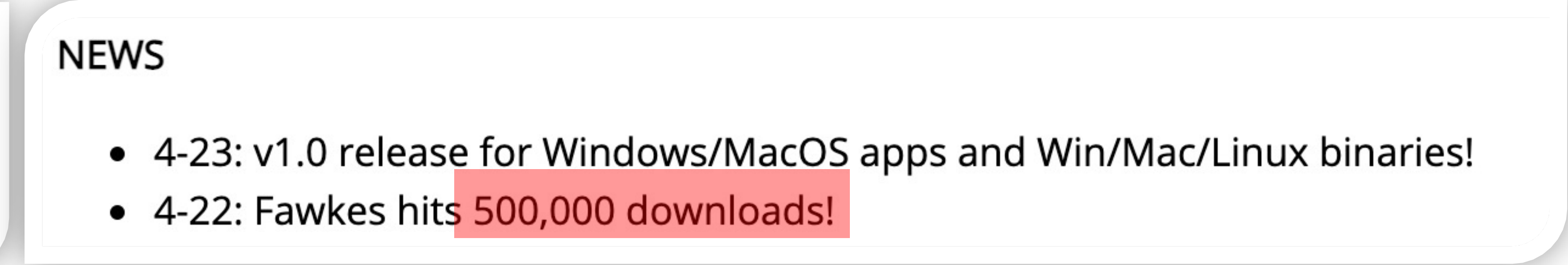
*This Tool Could Protect Your Photos From Facial Recognition*



sandlab.cs.uchicago.edu/fawkes

BSD-3-Clause License

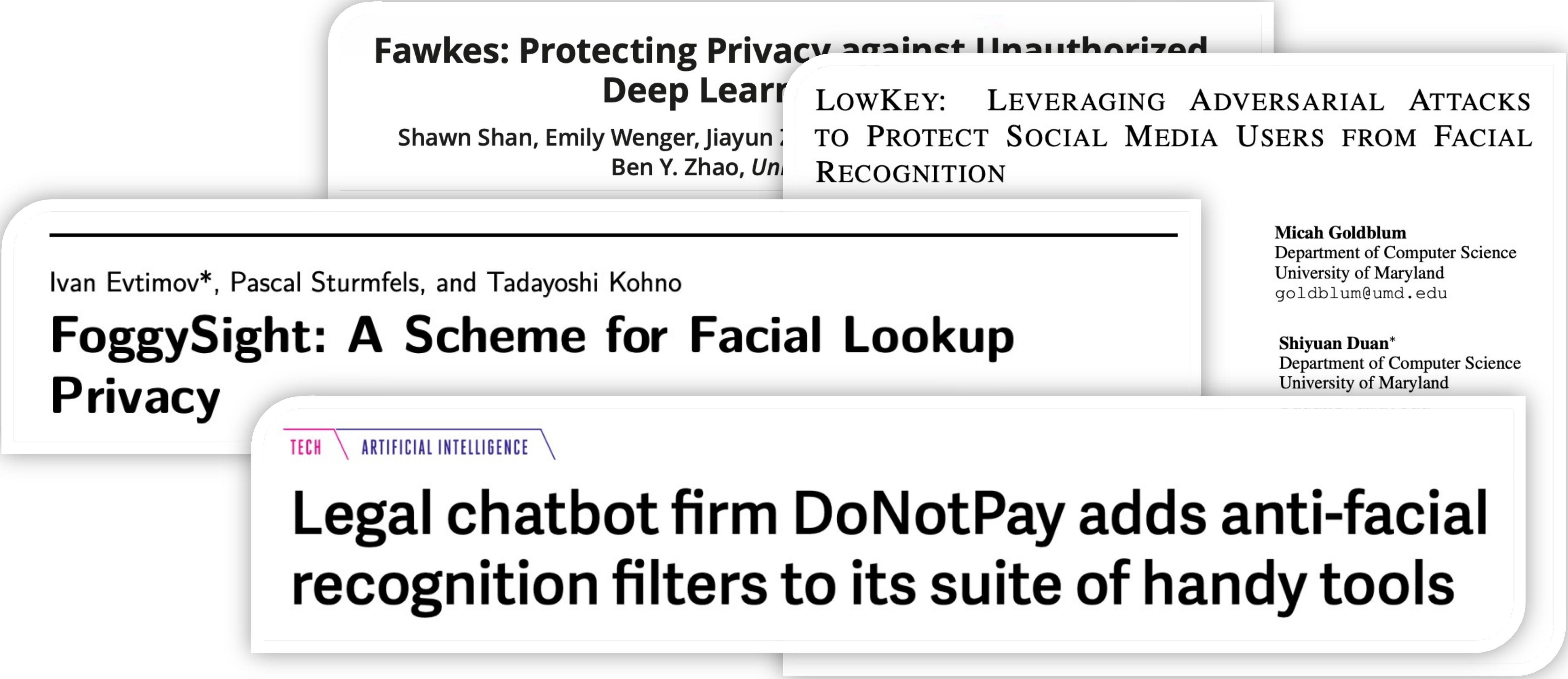
☆ 4.1k stars 402 forks



NEWS

- 4-23: v1.0 release for Windows/MacOS apps and Win/Mac/Linux binaries!
- 4-22: Fawkes hits 500,000 downloads!

# Image-perturbation tools promise to defeat facial recognition.



**Fawkes: Protecting Privacy against Unauthorized Deep Learn**

Shawn Shan, Emily Wenger, Jiayun
Ben Y. Zhao, *Uni*

LOWKEY: LEVERAGING ADVERSARIAL ATTACKS TO PROTECT SOCIAL MEDIA USERS FROM FACIAL RECOGNITION

**Micah Goldblum**
Department of Computer Science
University of Maryland
goldblum@umd.edu

**Shiyuan Duan**[*]
Department of Computer Science
University of Maryland

Ivan Evtimov*, Pascal Sturmfels, and Tadayoshi Kohno

**FoggySight: A Scheme for Facial Lookup Privacy**

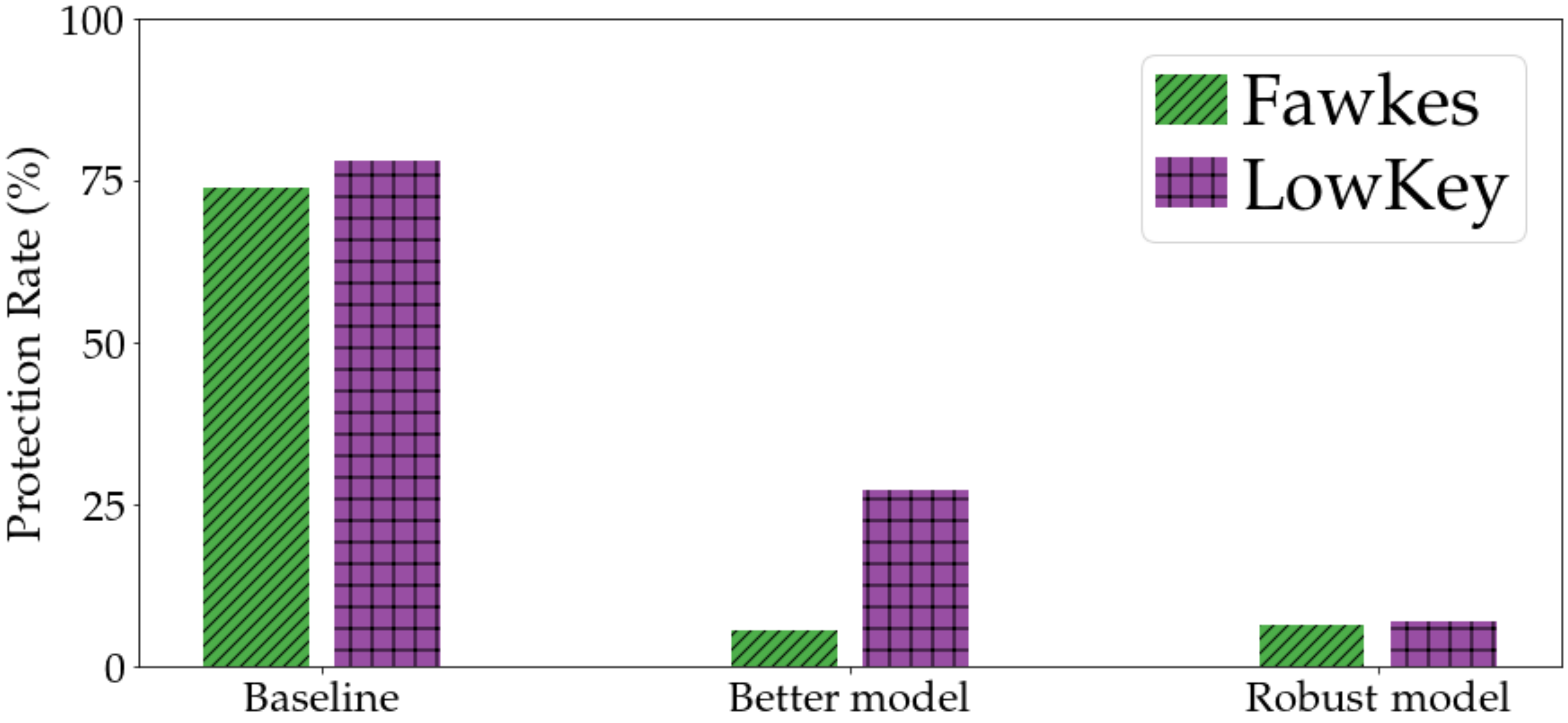
TECH ARTIFICIAL INTELLIGENCE

**Legal chatbot firm DoNotPay adds anti-facial recognition filters to its suite of handy tools**

These tools give a false sense of security!

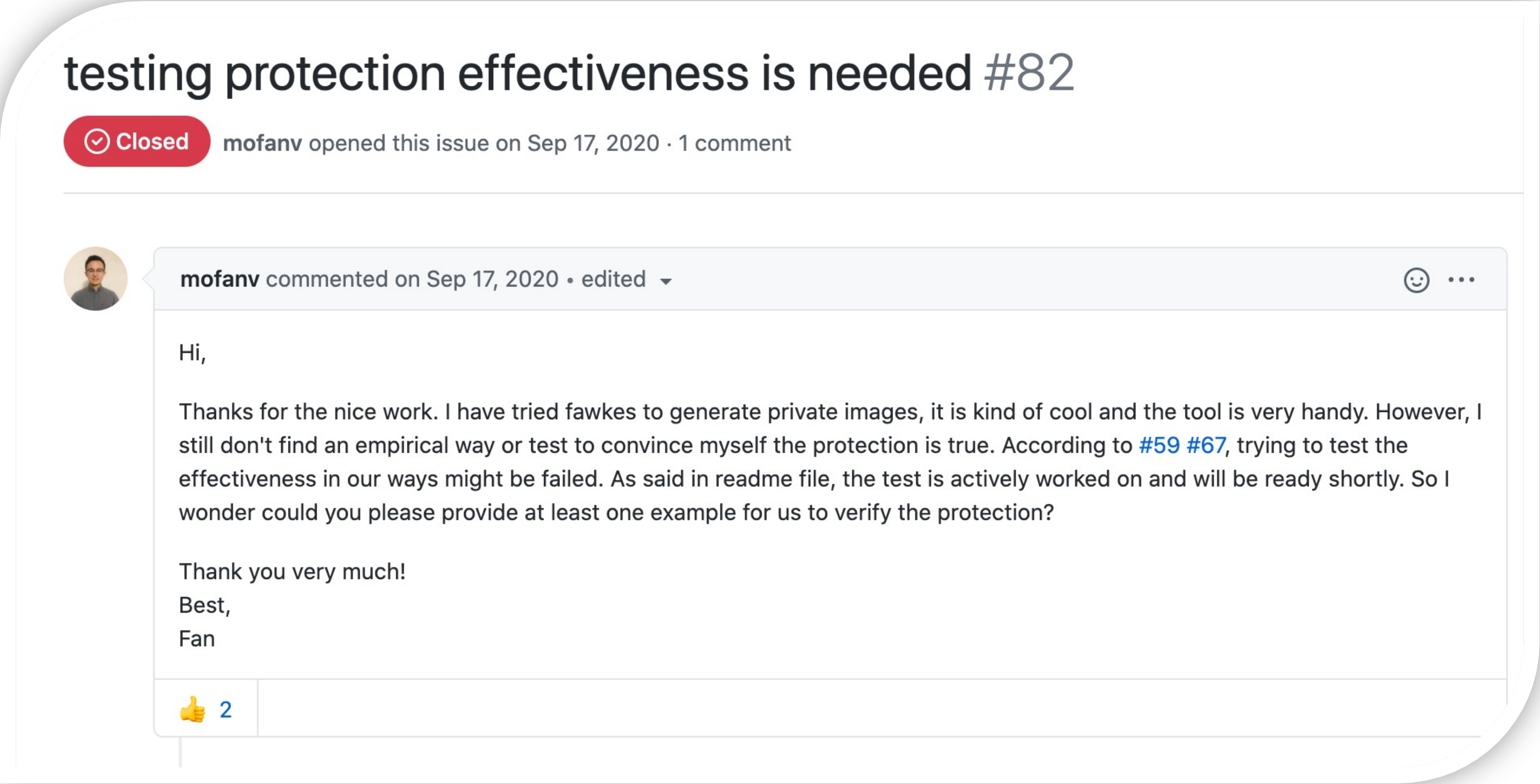
# These tools give a false sense of security!

## The attacks are easily defeated.



# These tools give a false sense of security!

## Users don't know if the attack worked…



testing protection effectiveness is needed #82

mofanv opened this issue on Sep 17, 2020 · 1 comment

mofanv commented on Sep 17, 2020 • edited

Hi,

Thanks for the nice work. I have tried fawkes to generate private images, it is kind of cool and the tool is very handy. However, I still don't find an empirical way or test to convince myself the protection is true. According to #59 #67, trying to test the effectiveness in our ways might be failed. As said in readme file, the test is actively worked on and will be ready shortly. So I wonder could you please provide at least one example for us to verify the protection?

Thank you very much!
Best,
Fan

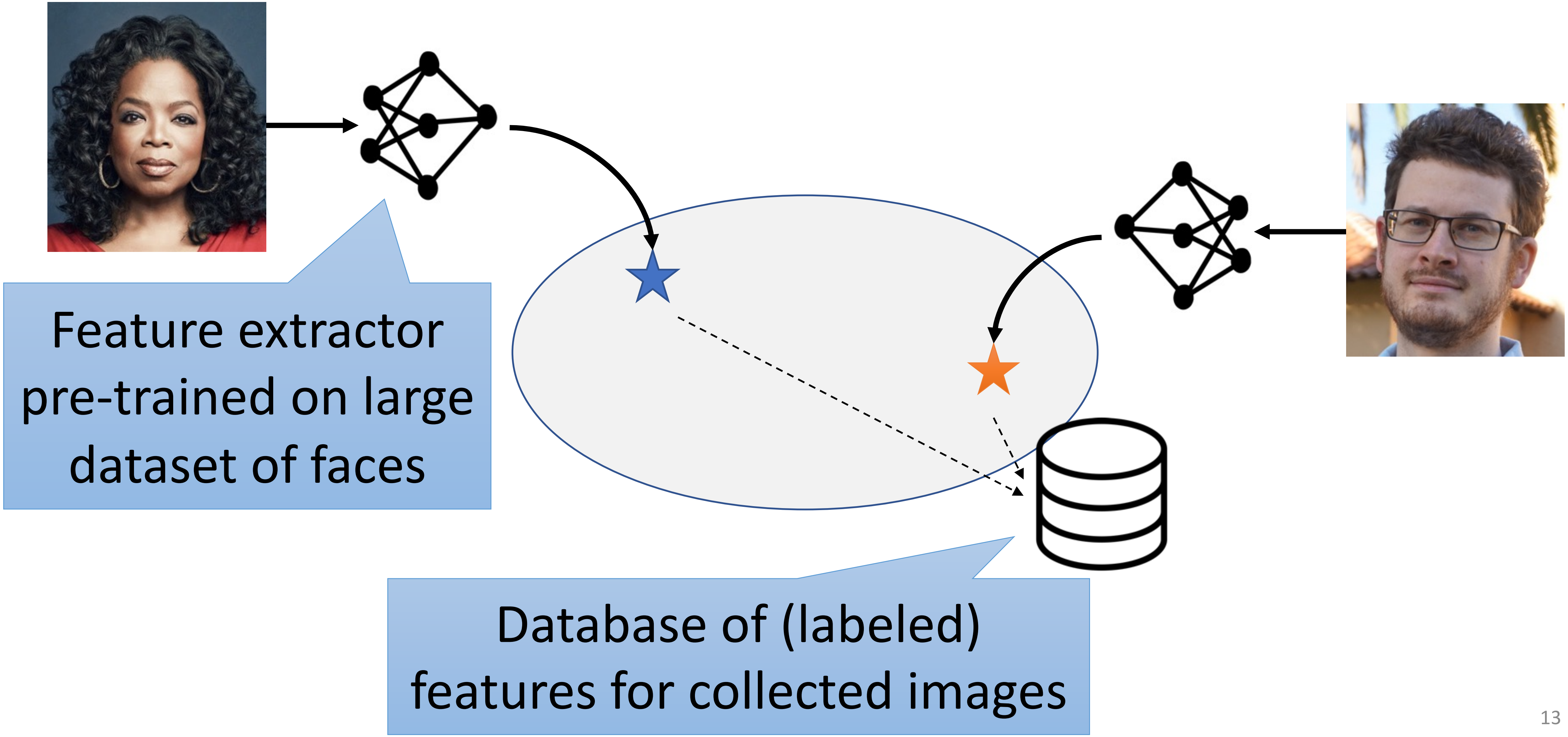
# This talk.

- Attacking facial recognition systems
- Misconceptions about adversarial examples
- Solutions?

# This talk.

- **Attacking facial recognition systems**
- Misconceptions about adversarial examples
- Solutions?

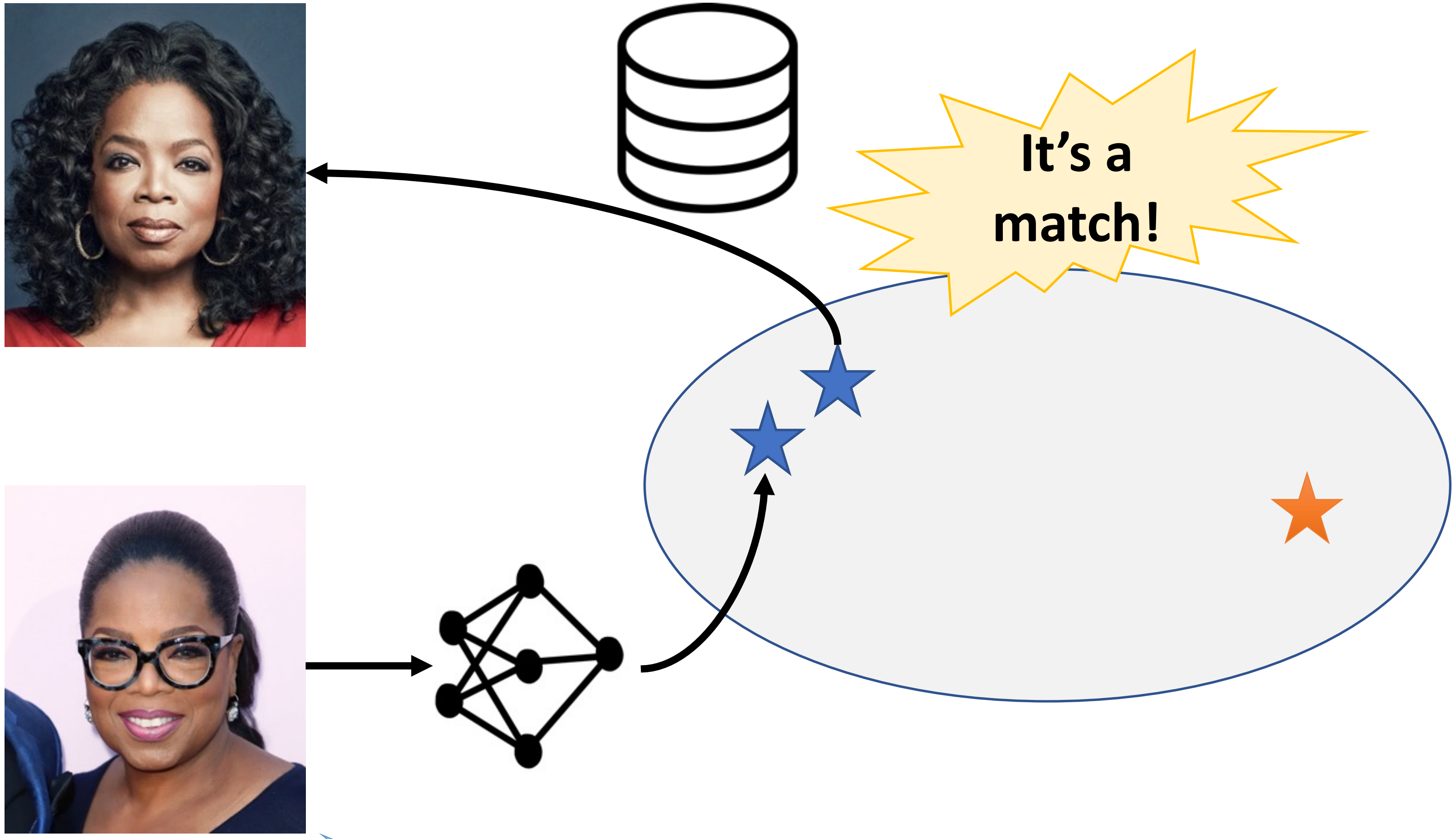
# Facial recognition with nearest neighbor search.



Feature extractor pre-trained on large dataset of faces

Database of (labeled) features for collected images

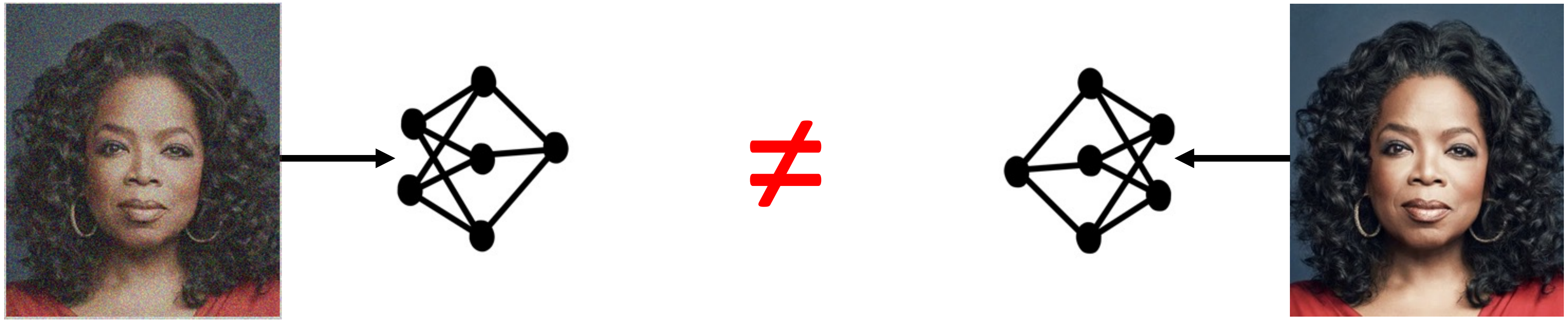
# Facial recognition with nearest neighbor search.



It's a match!

New image to match

# An attack: adversarial examples.

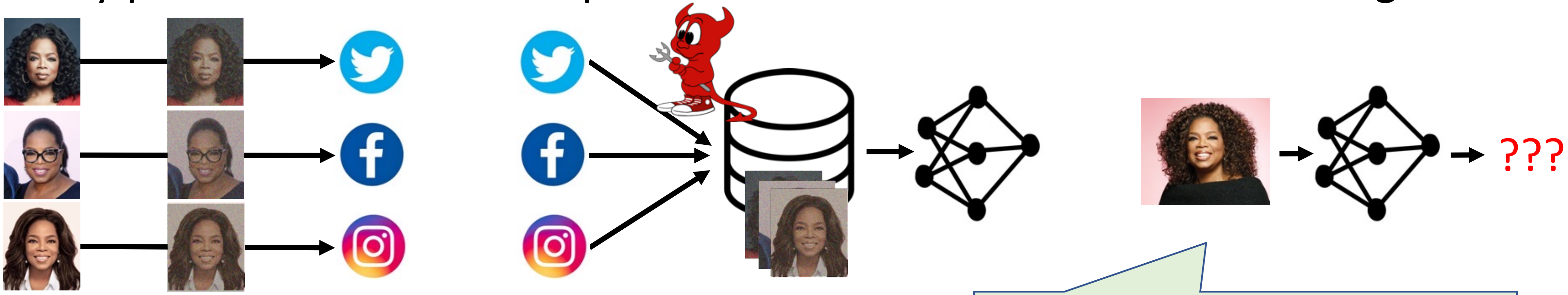


# Data poisoning with adversarial examples.

Users perturb pictures they post online

Online pictures are scraped to build a model

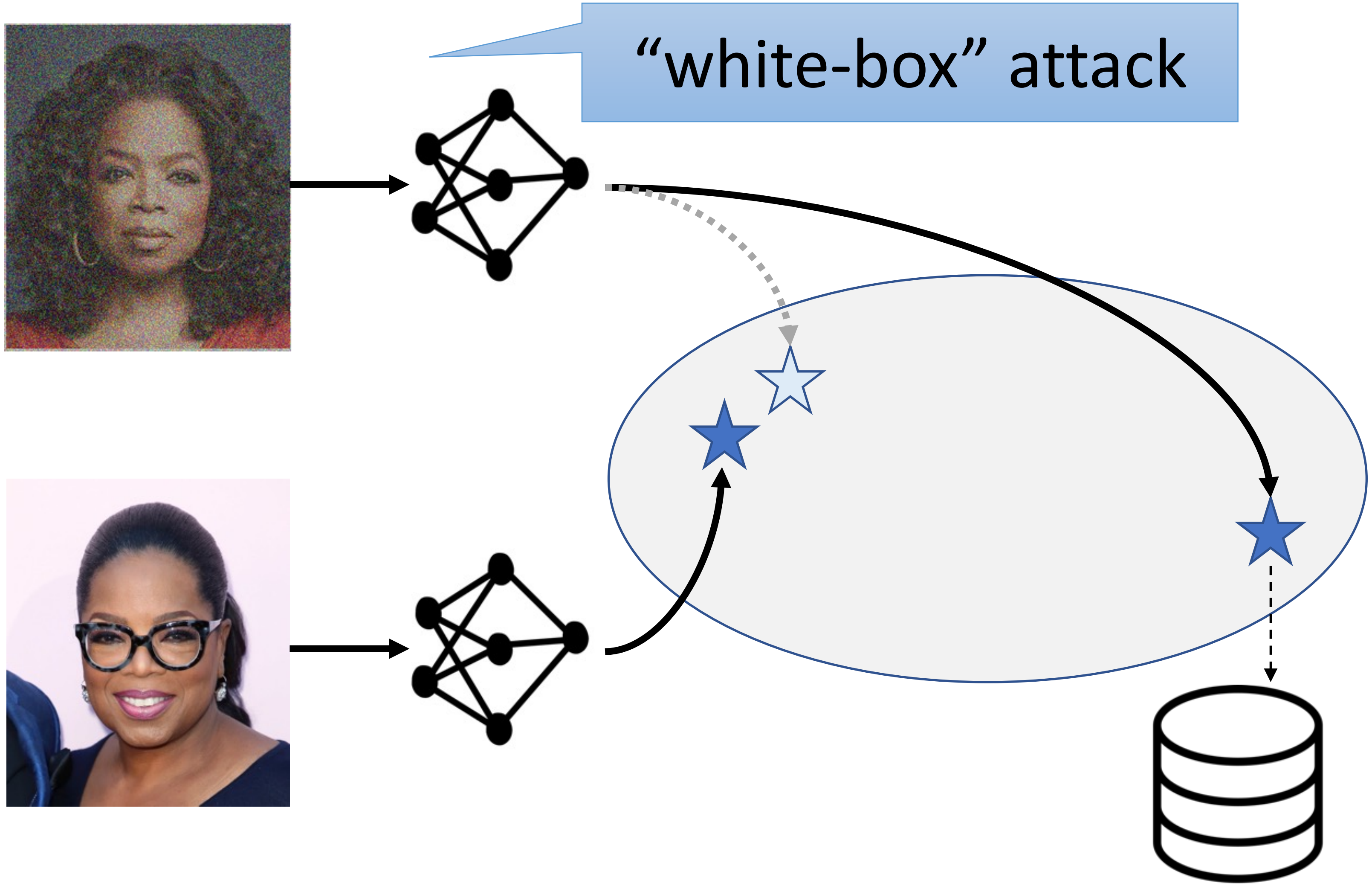
Unperturbed test pictures aren’t recognized



Users’ friends can still recognize the pictures

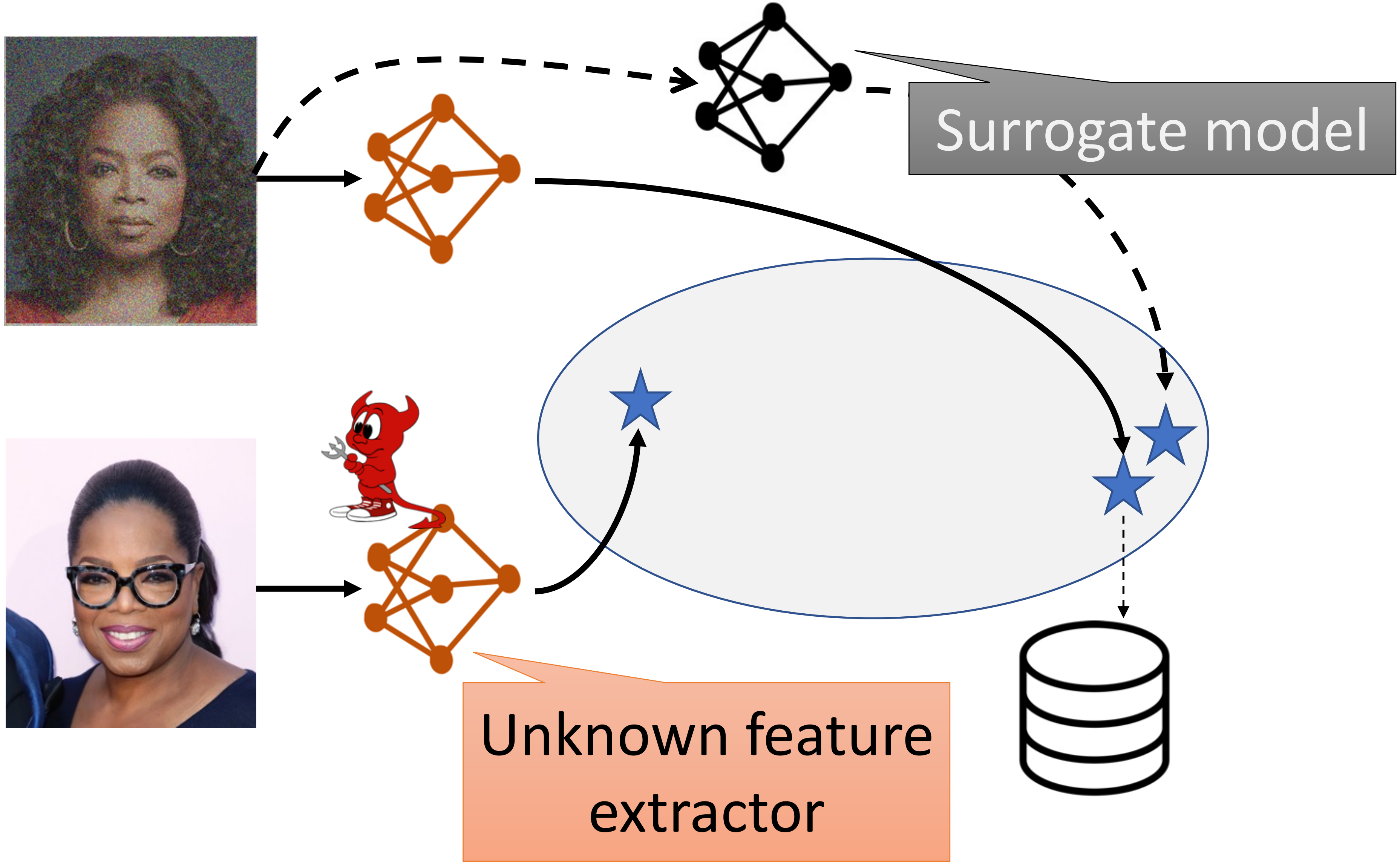
Unperturbed picture taken by the police, or a stalker, etc.

# Poisoning is easy if the extractor is *fixed & known*.



“white-box” attack

# The attack should transfer to unknown extractors.



Surrogate model

Unknown feature extractor

# This talk.

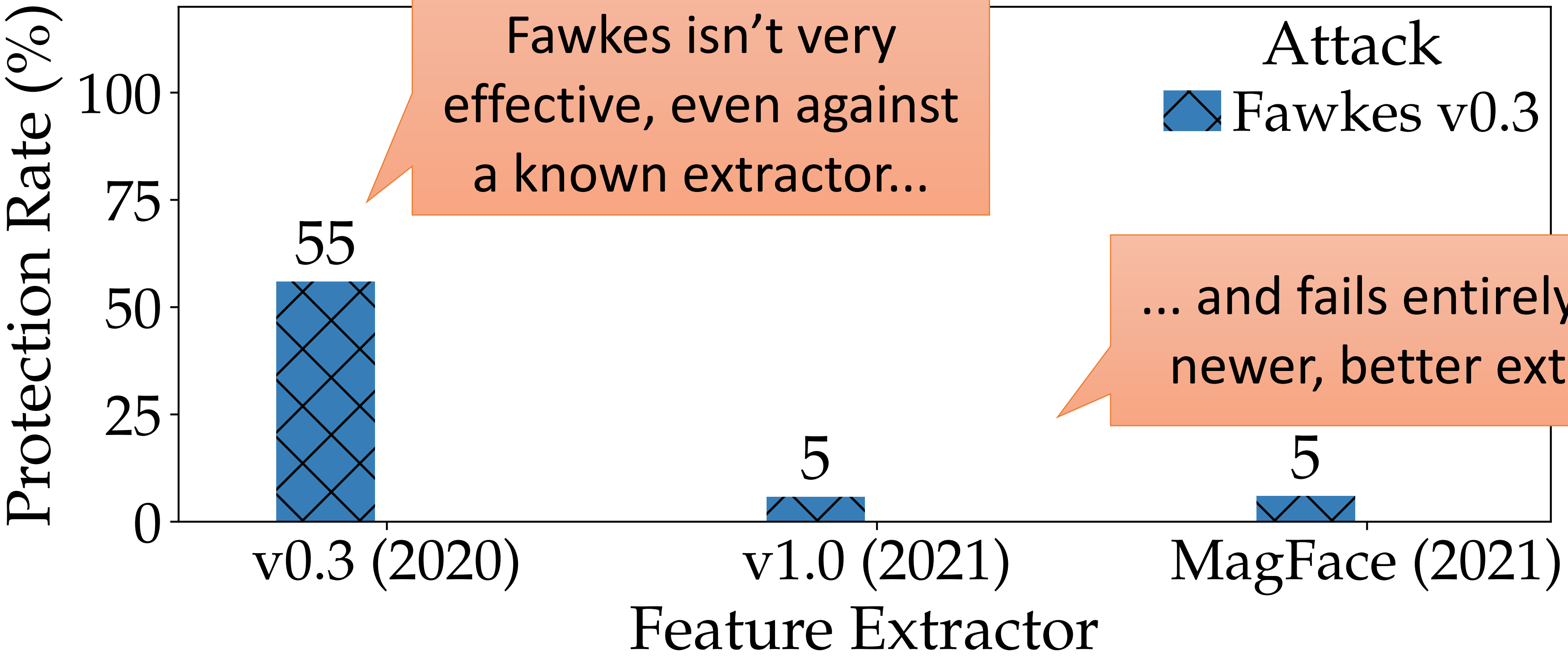
- Attacking facial recognition systems
- **Misconceptions about adversarial examples**
- Solutions?

# Misconception #1:

$$\underbrace{\forall \text{ models } \exists \text{ attack}}_{\text{this is empirically true (so far)}} \neq \exists \text{ attack } \forall \text{ models}$$

credit: Nicholas Carlini

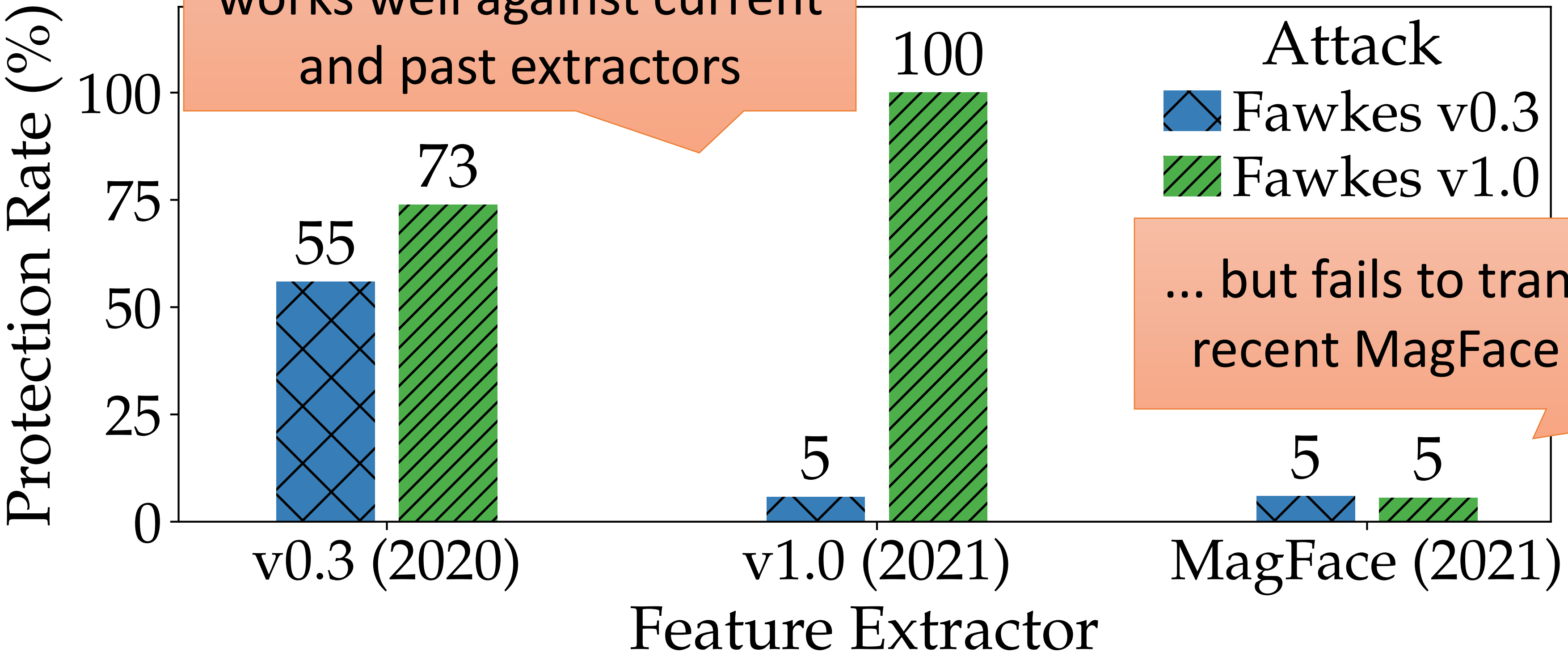
# Fawkes (v0.3) doesn’t transfer to today’s best models.



The 1st version of Fawkes isn’t very effective, even against a known extractor...

... and fails entirely against newer, better extractors

# Fawkes (v1.0) doesn't transfer to today's best models.

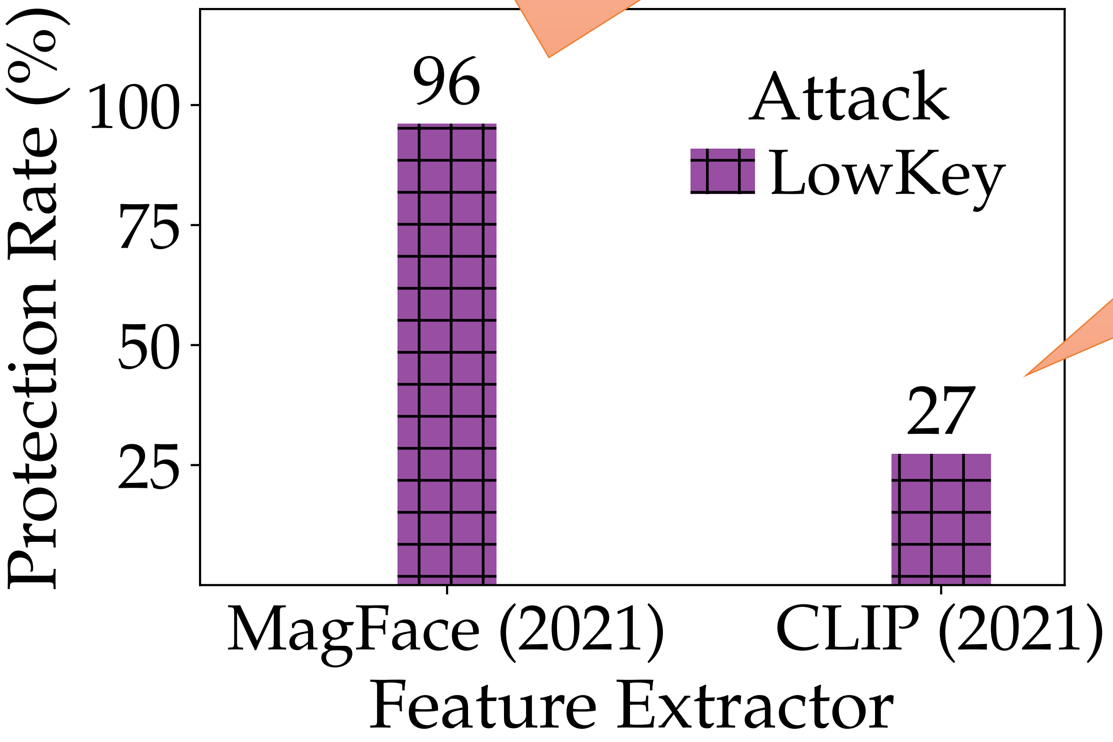


The new version of Fawkes works well against current and past extractors

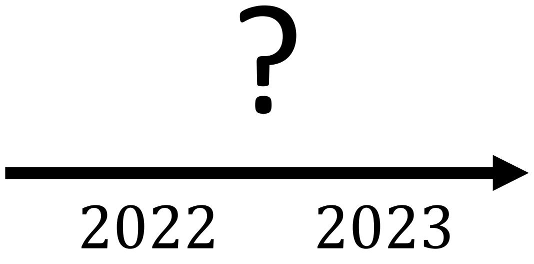
... but fails to transfer to the recent MagFace extractor

# LowKey transfers moderately to today's best models.

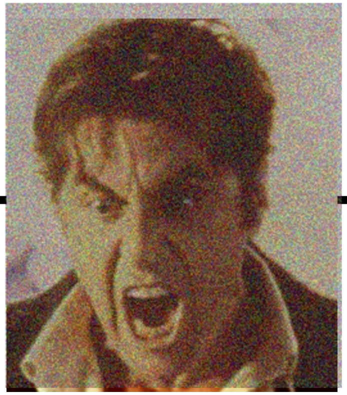
LowKey transfers very well against existing extractors...

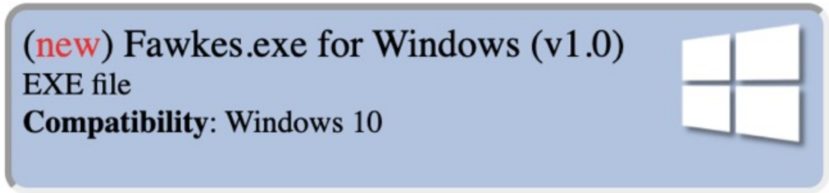


...but only moderately against a new and very different vision model

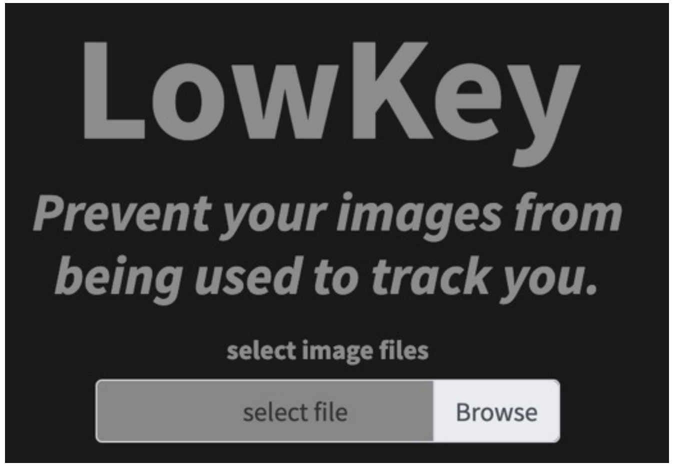


# What if the model trainer also uses the attack?





(new) Fawkes.exe for Windows (v1.0)
EXE file
**Compatibility**: Windows 10

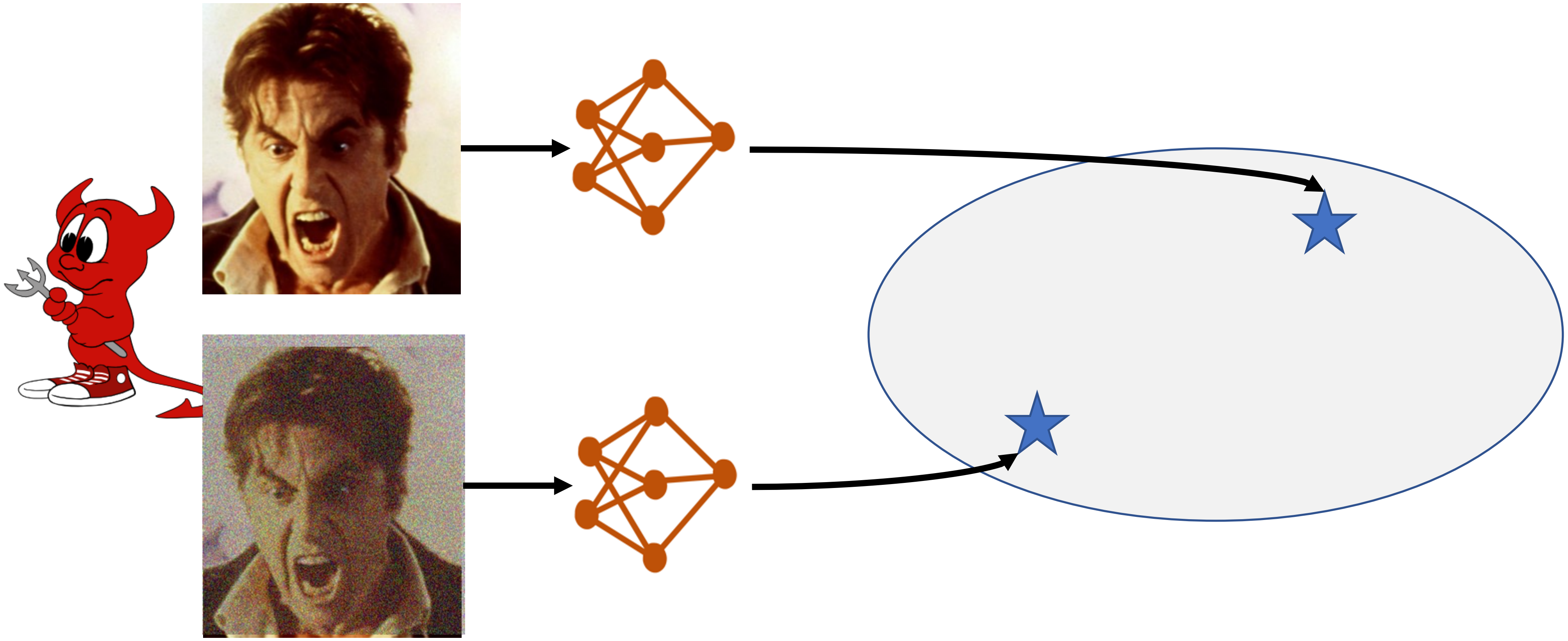


LowKey

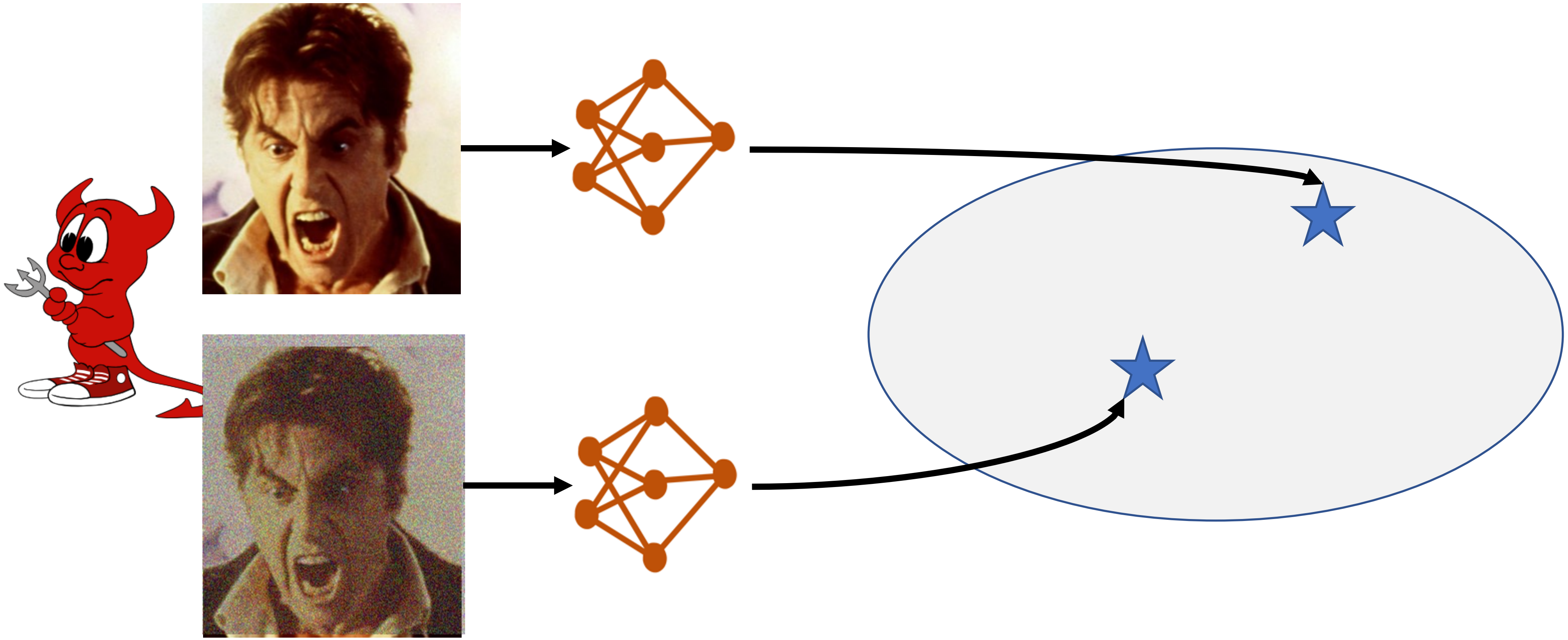
*Prevent your images from being used to track you.*

select image files

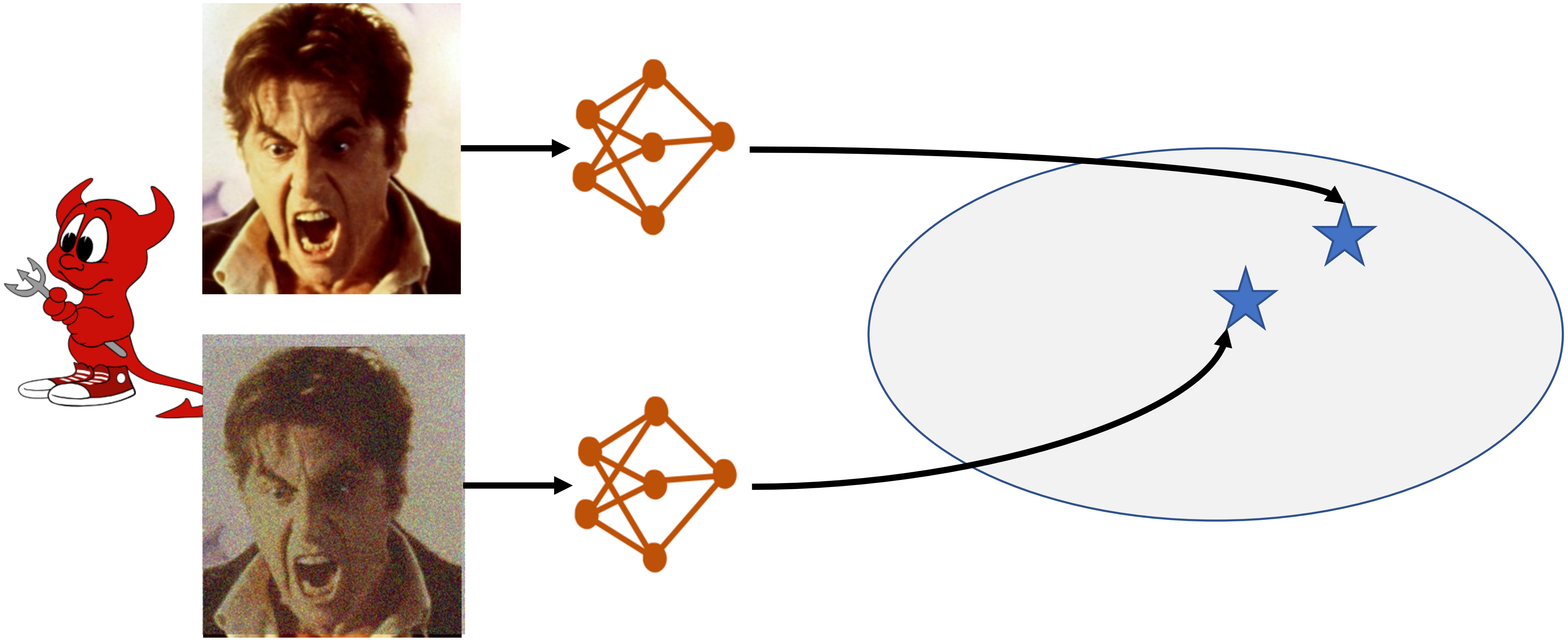
# Train a robust extractor on attack outputs.



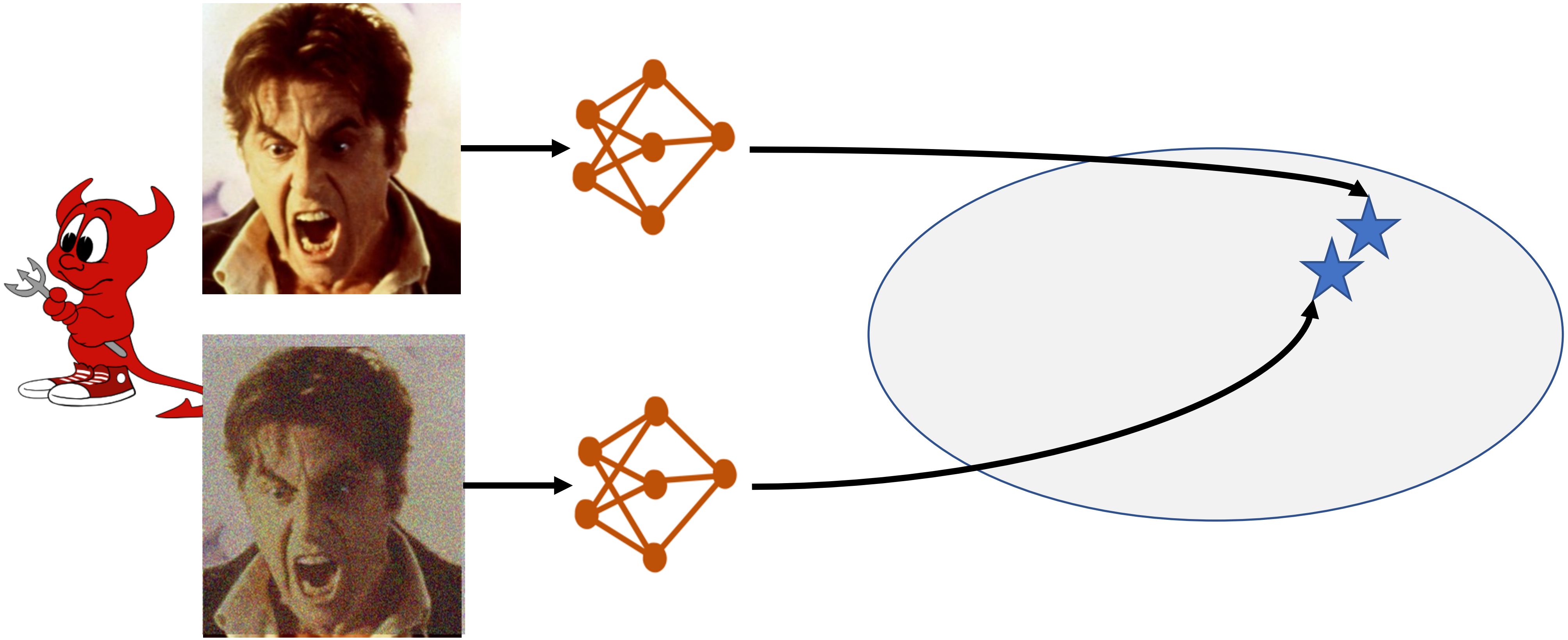
# Train a robust extractor on attack outputs.



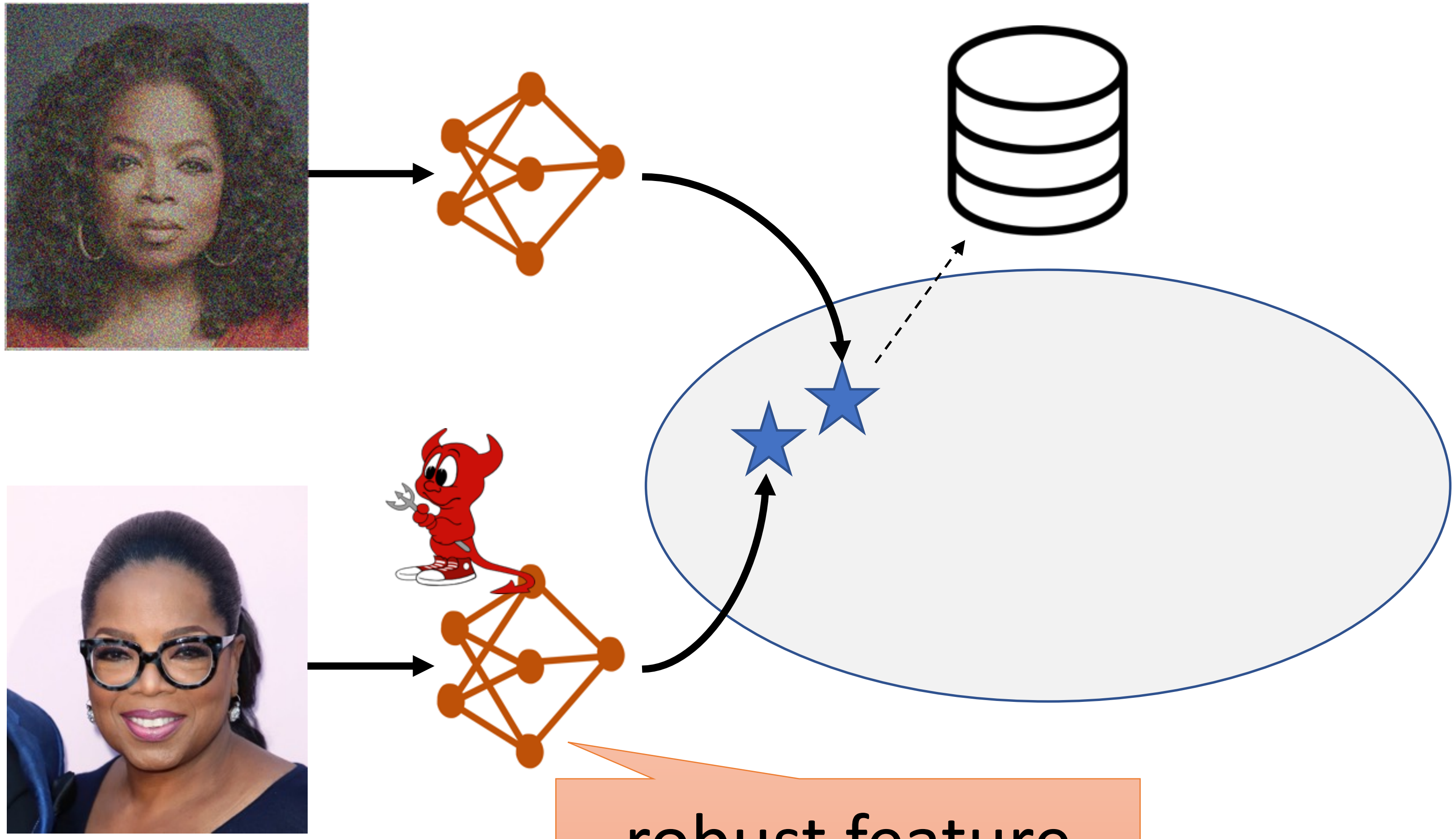
# Train a robust extractor on attack outputs.



# Train a robust extractor on attack outputs.

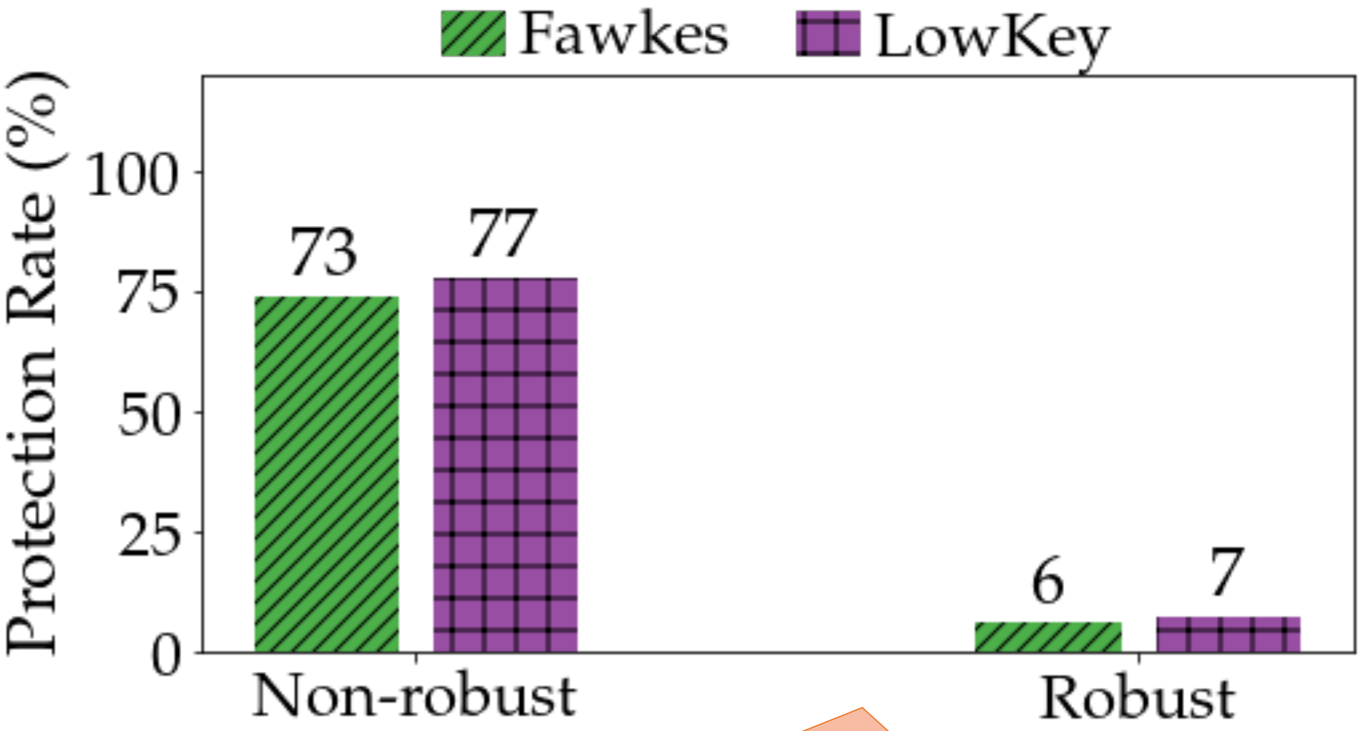


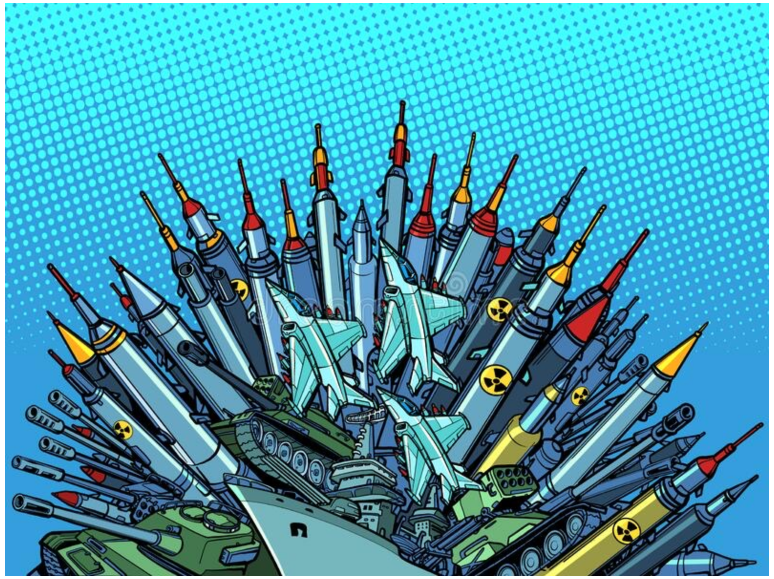
# The robust extractor resists poisoning attacks.



robust feature extractor

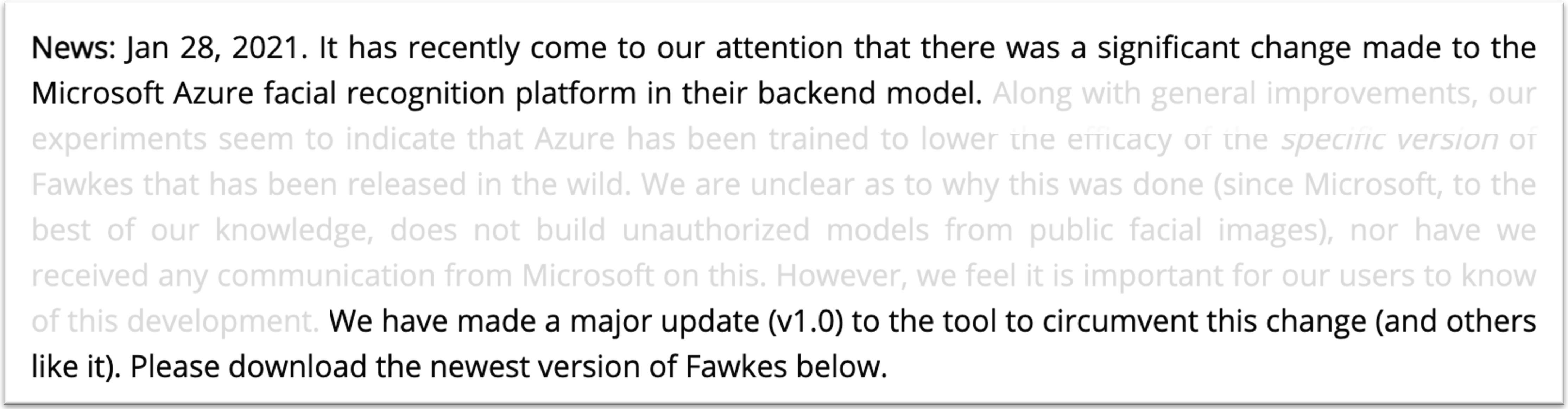
# The robust extractor resists poisoning attacks.





Isn’t this just the start of an arms-race?

# The arms race has already started.

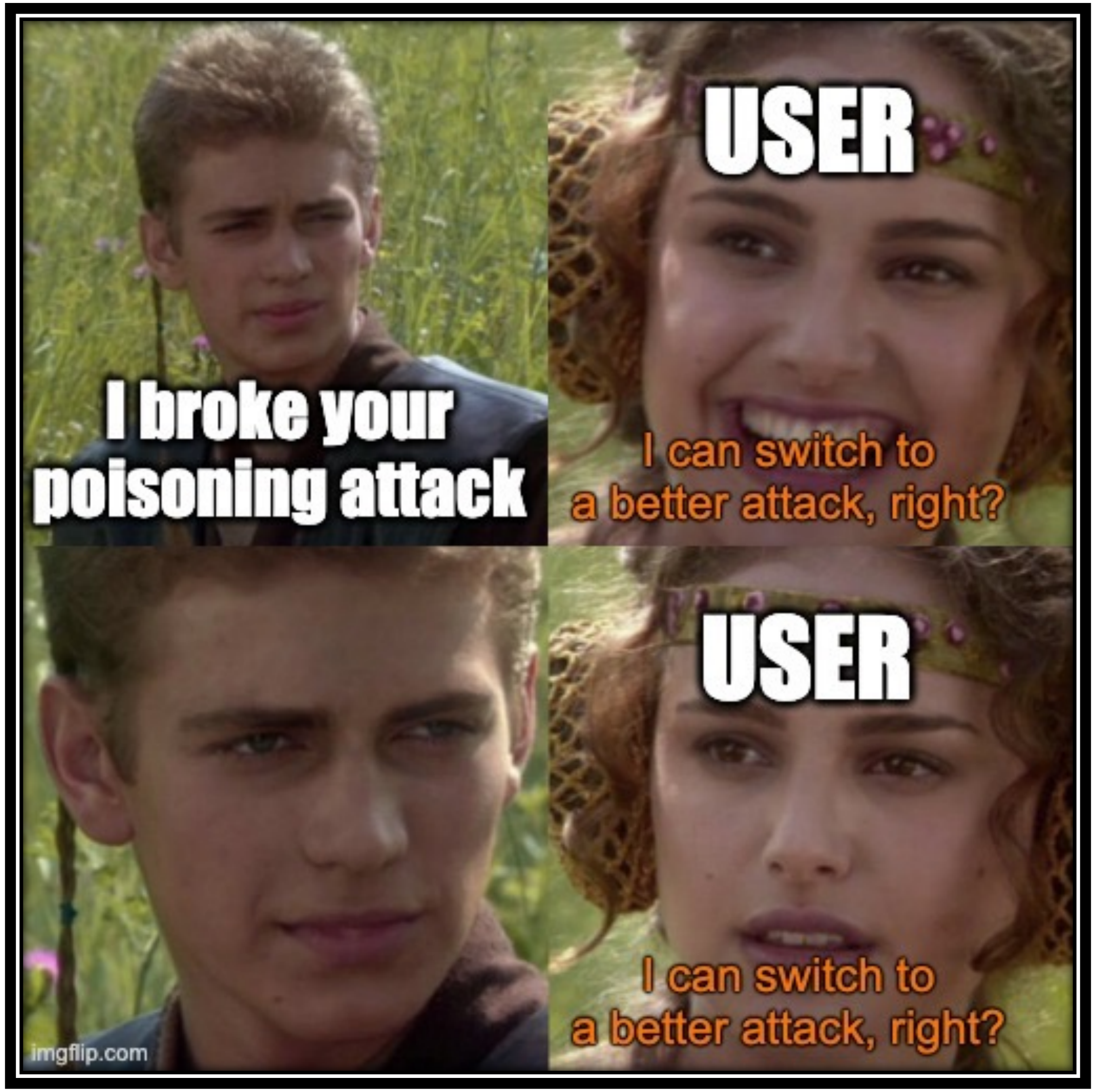


News: Jan 28, 2021. It has recently come to our attention that there was a significant change made to the Microsoft Azure facial recognition platform in their backend model. Along with general improvements, our experiments seem to indicate that Azure has been trained to lower the efficacy of the *specific version* of Fawkes that has been released in the wild. We are unclear as to why this was done (since Microsoft, to the best of our knowledge, does not build unauthorized models from public facial images), nor have we received any communication from Microsoft on this. However, we feel it is important for our users to know of this development. We have made a major update (v1.0) to the tool to circumvent this change (and others like it). Please download the newest version of Fawkes below.

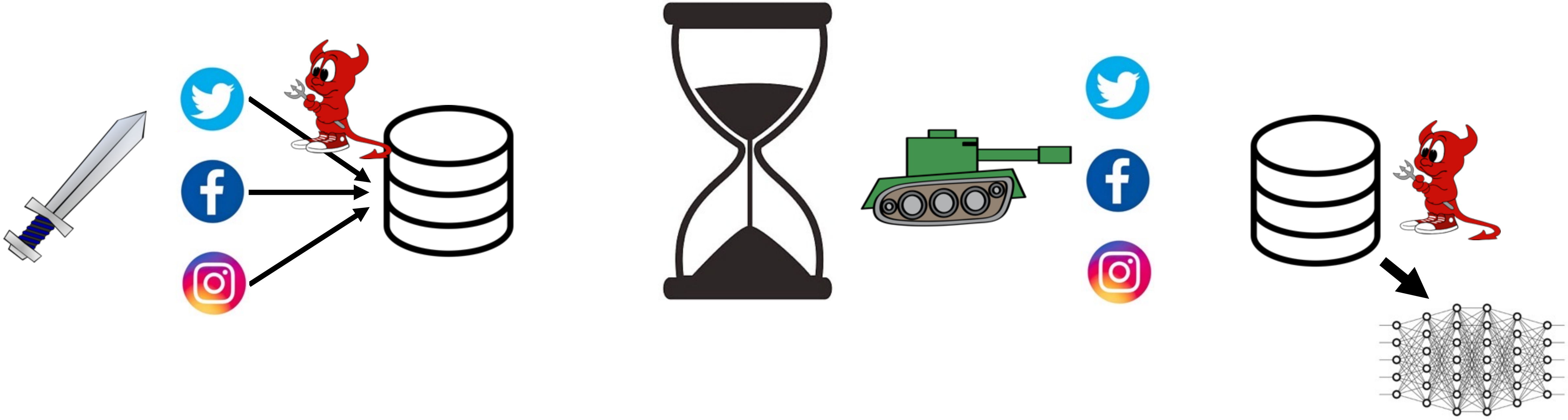
https://sandlab.cs.uchicago.edu/fawkes/

# Misconception #2:

Adversarial ML doesn’t always admit an arms-race


I broke your poisoning attack
USER
I can switch to a better attack, right?
USER
I can switch to a better attack, right?
imgflip.com

# New models can be applied *retroactively*.



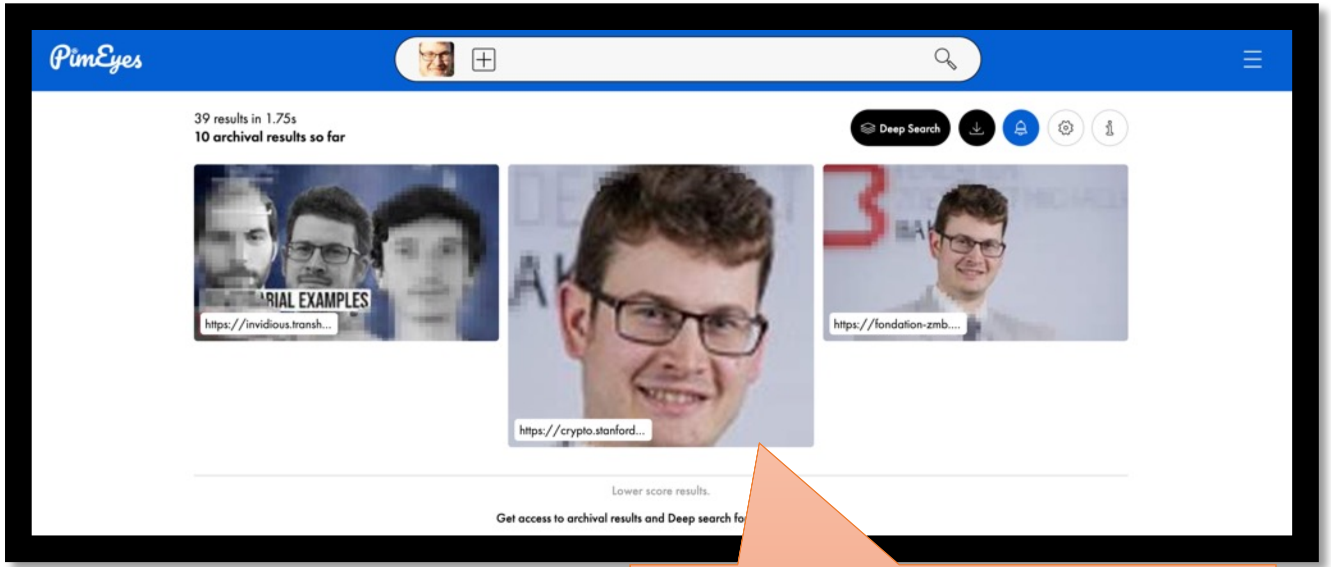
Model trainer scrapes pictures produced with weak attack

Users switch to stronger attacks, but new model can be trained solely on data collected in the past

# (Biometric) privacy does not admit an arms race.

- Facial features cannot be (easily or quickly) changed

- **You cannot reclaim your privacy once you've lost it!**



39 results in 1.75s

10 archival results so far

https://invidious.transh...

https://crypto.stanford...

https://fondation-zmb....

Lower score results.

Get access to archival results and Deep search fo

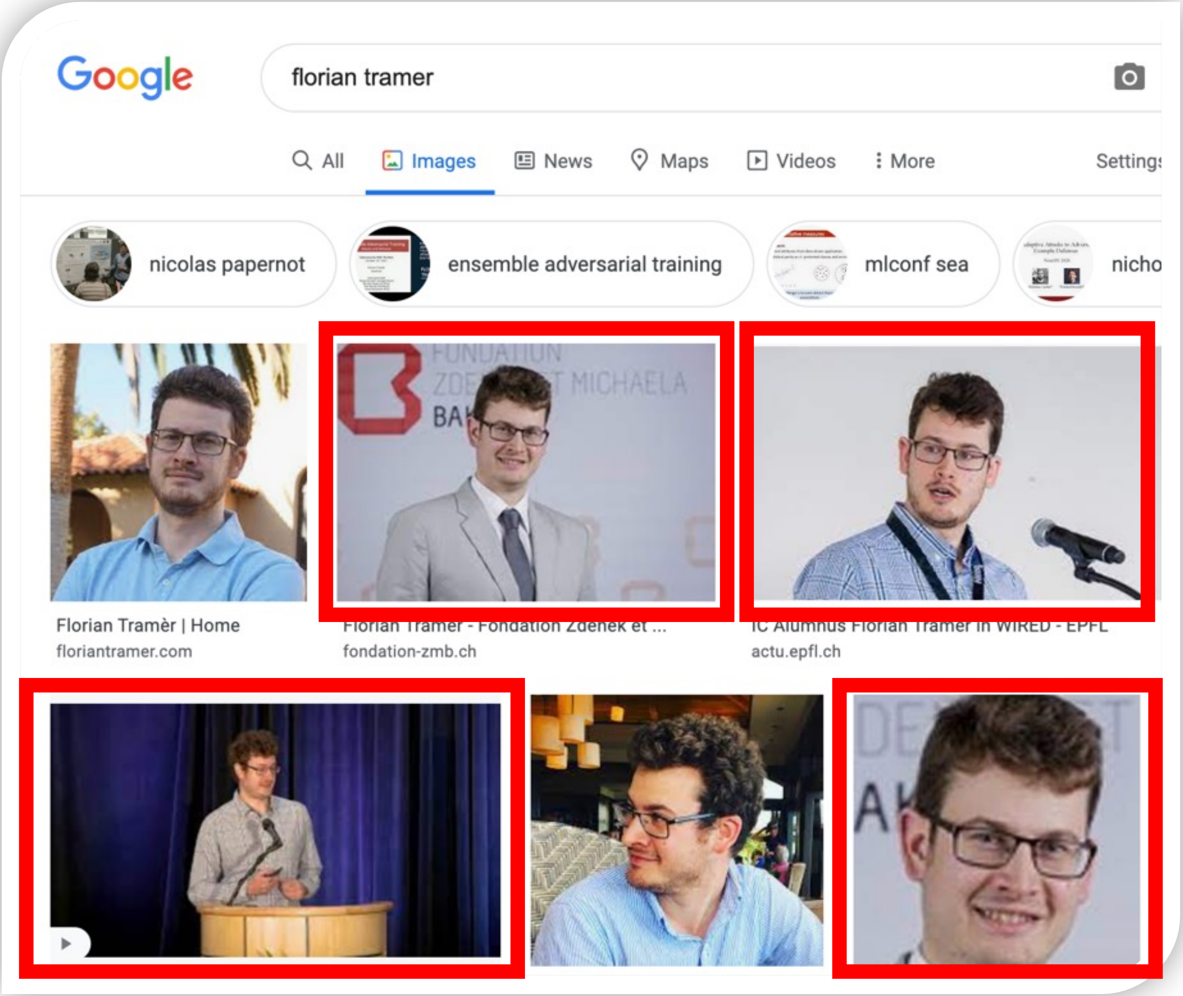
~6 years ago

# This talk.

- Attacking facial recognition systems
- Misconceptions about adversarial examples
- **Solutions?**

# Solution 1: Don’t post pictures online.

# Solution 1: Don’t post pictures online.
# It’s already too late.



florian tramer

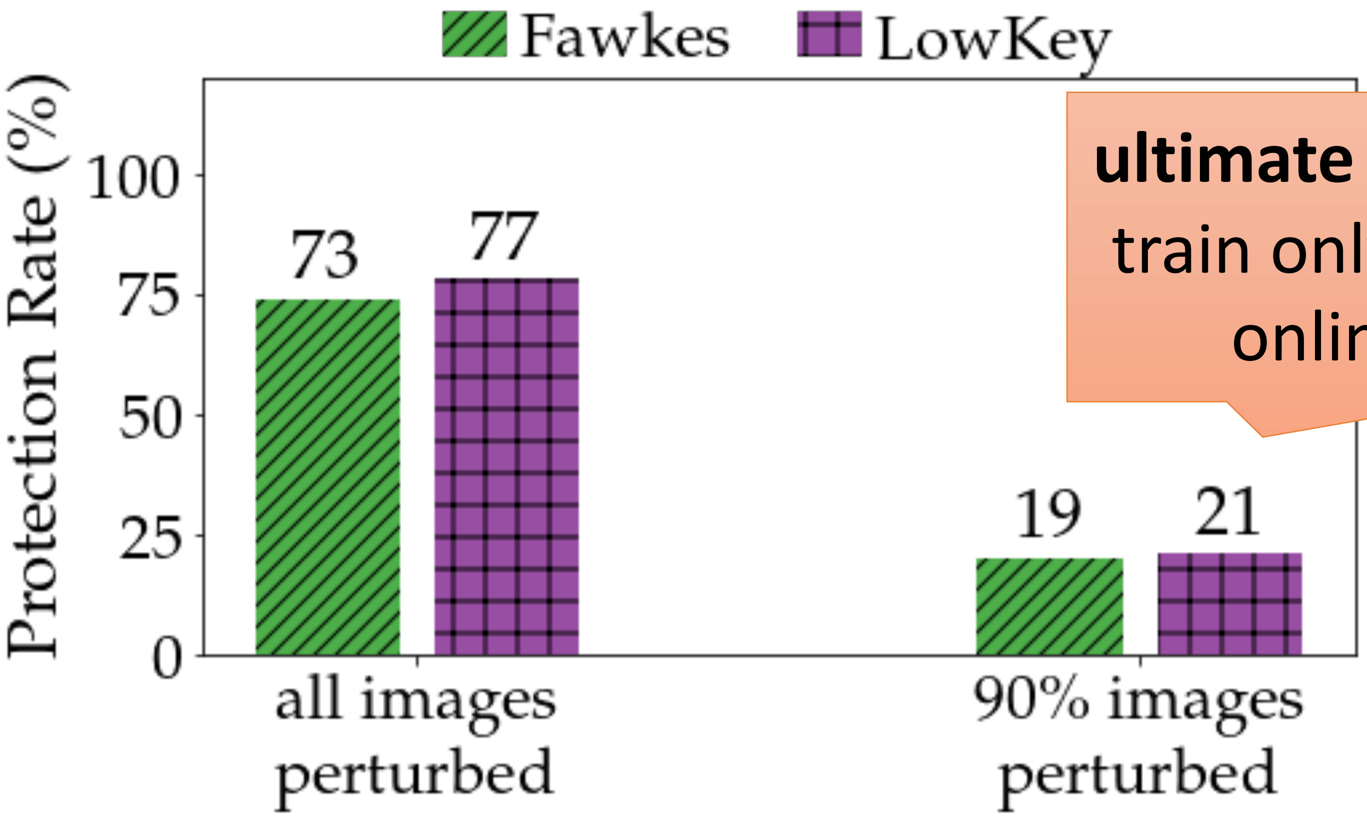
Florian Tramèr | Home
floriantramer.com

Florian Tramer - Fondation Zdenek et ...
fondation-zmb.ch

IC Alumnus Florian Tramer in WIRED - EPFL
actu.epfl.ch

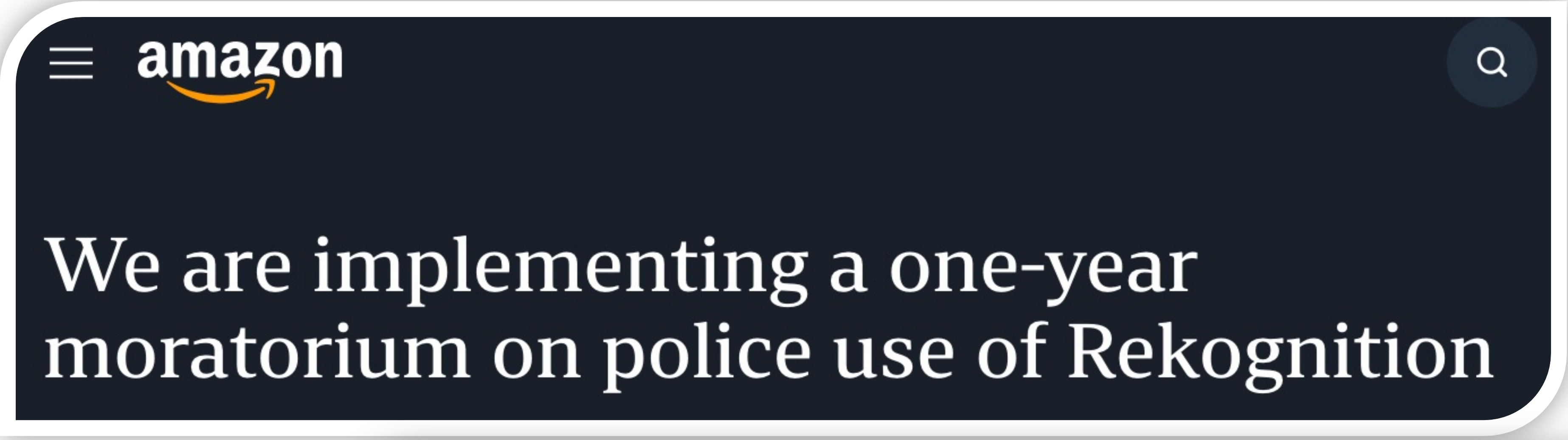
I didn’t take these pictures or upload them!

# Solution 1: Don’t post pictures online.
# It’s already too late.

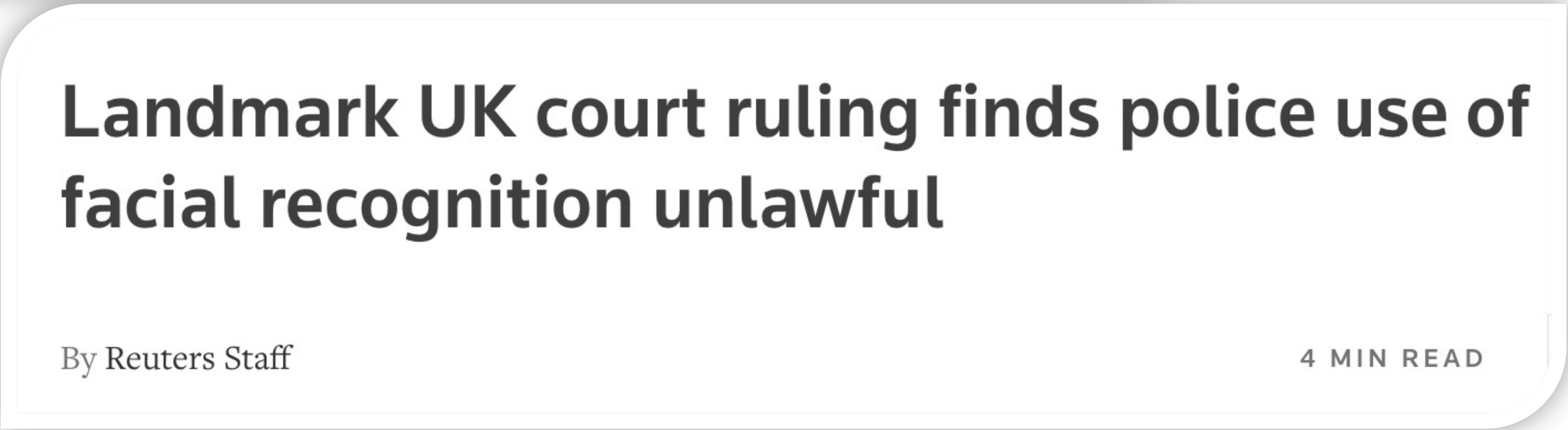


**ultimate retroactive strategy:** train only on pictures posted online before 2020...

# Solution 2: Legislation & policy



We are implementing a one-year moratorium on police use of Rekognition



Landmark UK court ruling finds police use of facial recognition unlawful

By Reuters Staff

4 MIN READ

# Take-Aways

- Threat models matter:
  - no single attack works against *all future* models
  - biometric privacy does **not** admit an arms race

- Be careful what you can promise users

**Thank you!**