UNIVERSITY OF WATERLOO

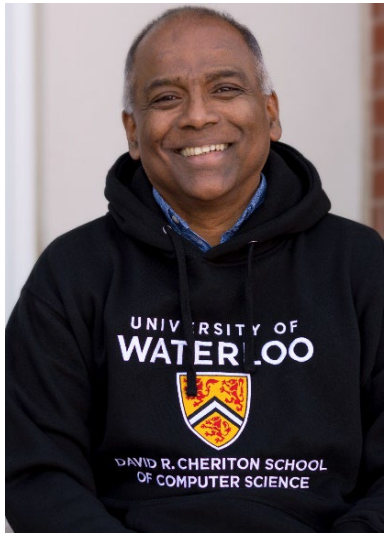
# Confidence in AI

## Can we trust AI-based systems?

*N. Asokan*
https://asokan.org/asokan/
@nasokan

*(Joint work with Buse Gul Atli, Sebastian Szyller, Mika Juuti, Jian Liu, Rui Zhang, Vasisht Duddu, Asim Waheed, and Samuel Marchal)*

# Who am I?




**Executive Director, Cybersecurity and Privacy Institute (CPI)**
**Professor of Computer Science, University of Waterloo**

**Fellow of the Royal Society of Canada (2023), IEEE Fellow (2017), ACM Fellow (2019)**

***Previously*: Professor, Aalto University (2013-2019), Nokia** (14 y; built up Nokia security research team), **IBM Research** (3 y)

**Industry collaborations: Private-AI Institute, ICRI-CARS, Google Awards**

**https://asokan.org/asokan/ for more background**

# University of Waterloo

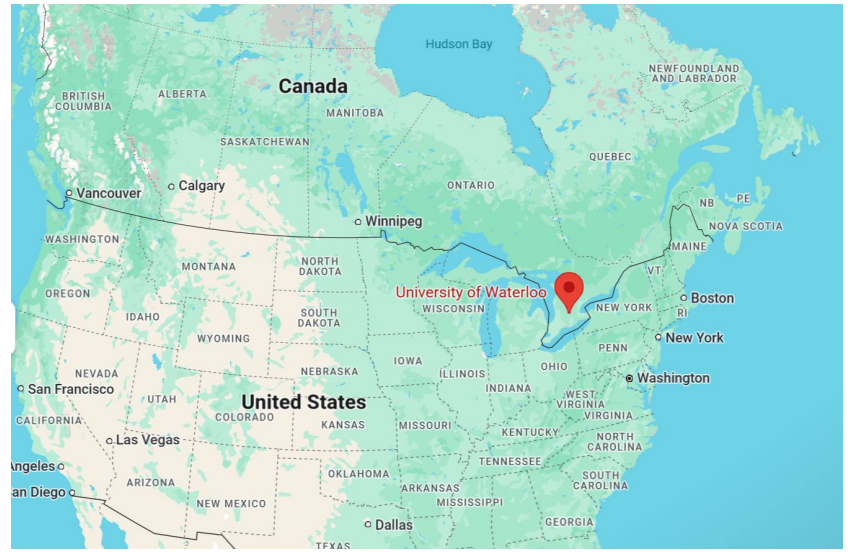
**#1 in cybersecurity/privacy technology research in Canada; top-15 in the world**
**https://csrankings.org/**

**100+ professors in computer science**

**60+ professors (across the university) working on different aspects of cybersecurity/privacy**

**#1 Engineering school in Canada**

UNIVERSITY OF WATERLOO




https://uwaterloo.ca/bioinformatics-group/about

# My research interests

**Systems Security and Privacy**

## AI and Security/Privacy

- How to use AI to improve security/privacy solutions
- How to improve security/privacy of AI-based systems

## Platform security

- How to use hardware assistance to secure software?

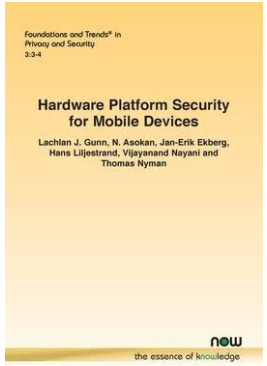
https://ssg-research.github.io/

# Platform security research

https://ssg-research.github.io/platsec/

## Hardware assisted trusted execution environments (TEEs)


CCS 2019 keynote[1] https://youtu.be/hHYoGn5PSI4




2022 book https://ssg.aalto.fi/publications/hardware-platform-security-for-mobile-devices/

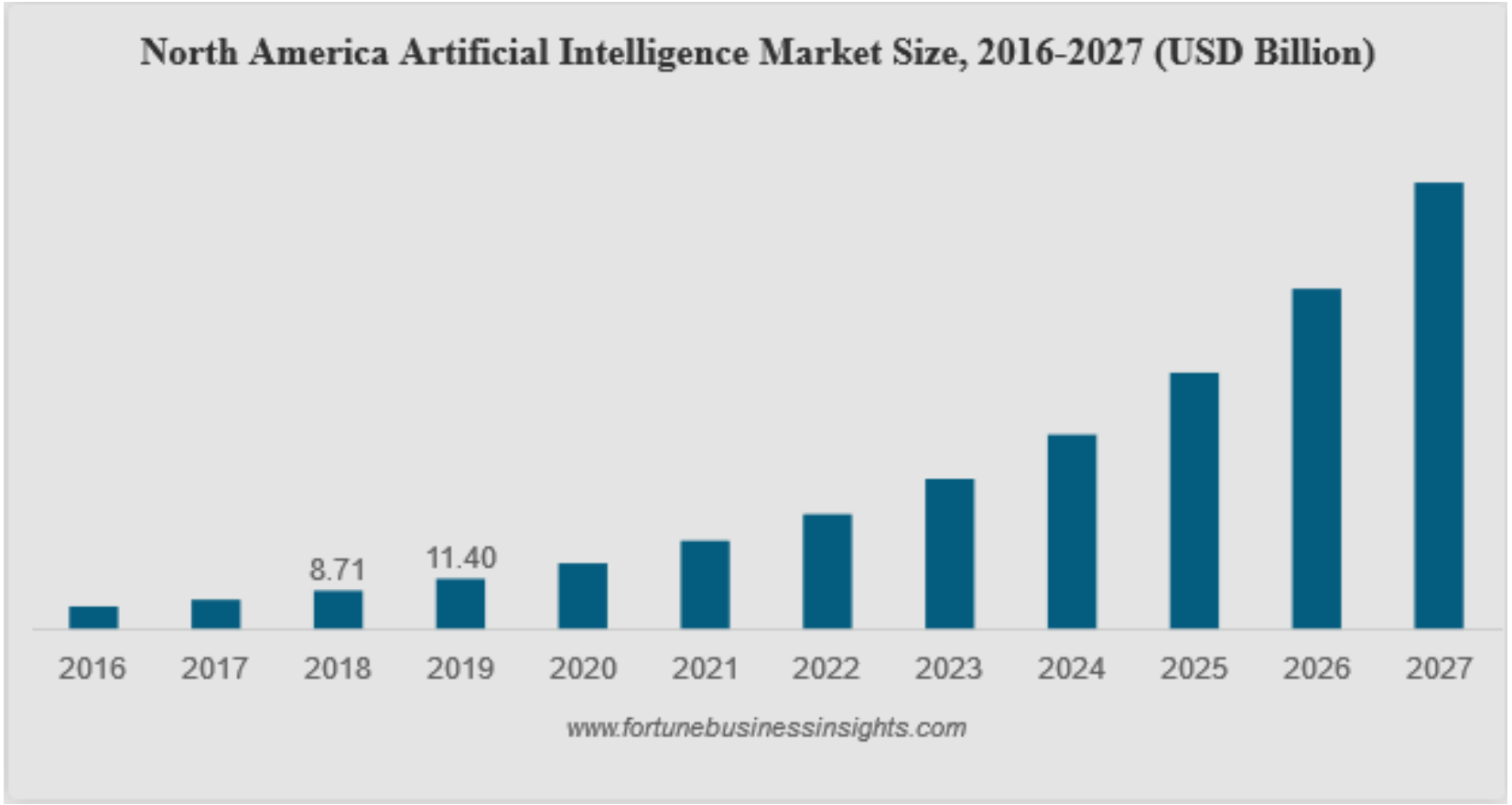
## Novel hardware security mechanisms

- HardScope (DAC 2019, https://arxiv.org/abs/1705.10295) , BliMe (NDSS 2024, HOST 2024, https://ssg-research.github.io/platsec/blime)

## Novel uses of deployed hardware security mechanisms

- PACStack (Usenix SEC 2021, https://arxiv.org/abs/1905.10242) and PARTS (Usenix SEC 2019, https://arxiv.org/abs/1811.09189), Deterministic MTE tagging (https://arxiv.org/abs/2204.03781)

# AI will be pervasive




https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114

Forbes

7,109 views | Oct 18, 2019, 01:56pm EDT

# How Artifical Intelligence Is Advancing Precision Medicine

**Nicole Martin** Former Contributor ⓘ

AI & Big Data

*I write about digital marketing, data and privacy concerns.*

https://www.forbes.com/sites/nicolemartin1/2019/10/18/how-artifical-intelligence-is-advancing-precision-medicine/#2f720a79a4d5

MOTHERBOARD

TECH BY VICE

# Dozens of Cities Have Secretly Experimented With Predictive Policing Softw

Documents obtained by Motherboa
requests verify previously unconfir
with predictive policing company P

By Caroline Haskins

https://www.vice.com/en_us/article/d3m
experimented-with-predictive-policing-s

Forbes

5,705 views | Oct 31, 2019, 02:42pm EDT

# How AI Is Uprooting Recruiting

**Falon Fatemi** Contributor ⓘ

Entrepreneurs

https://www.vice.com/en_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software

PART OF A ZDNET SPECIAL FEATURE: CYBERSECURITY: LET'S GET TACTICAL

# AI is changing everything about cybersecurity, for better and for worse. Here's what you need to know

Artificial intelligence and machine learning tools could go a long way to helping to fight cybercrime. But these technologies aren't a silver bullet, and could also be exploited by malicious hackers.

https://www.zdnet.com/article/ai-is-changing-everything-about-cybersecurity-for-better-and-for-worse-heres-what-you-need-to-know/

# Challenges in making AI trustworthy

Security concerns

Privacy concerns

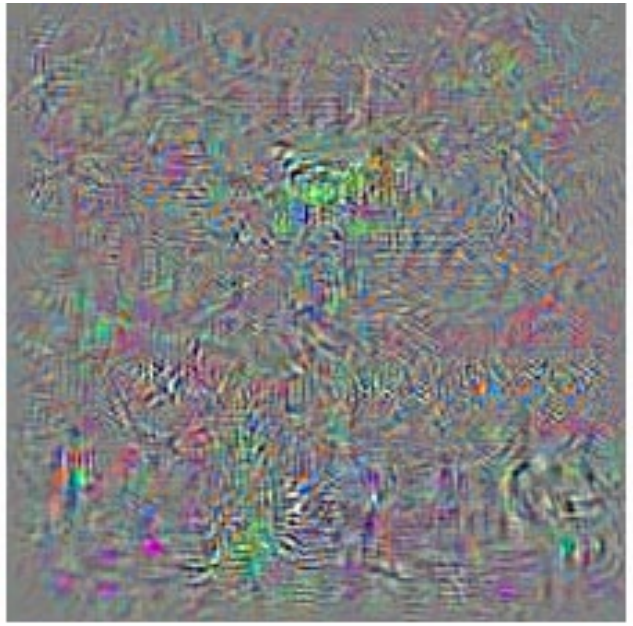
[Other concerns: fairness, explainability, alignment]

# Evading machine learning models


**Which class is this?**

**School bus**

**+ 0.1 ·**



**=**


**Which class is this?**

**Ostrich**

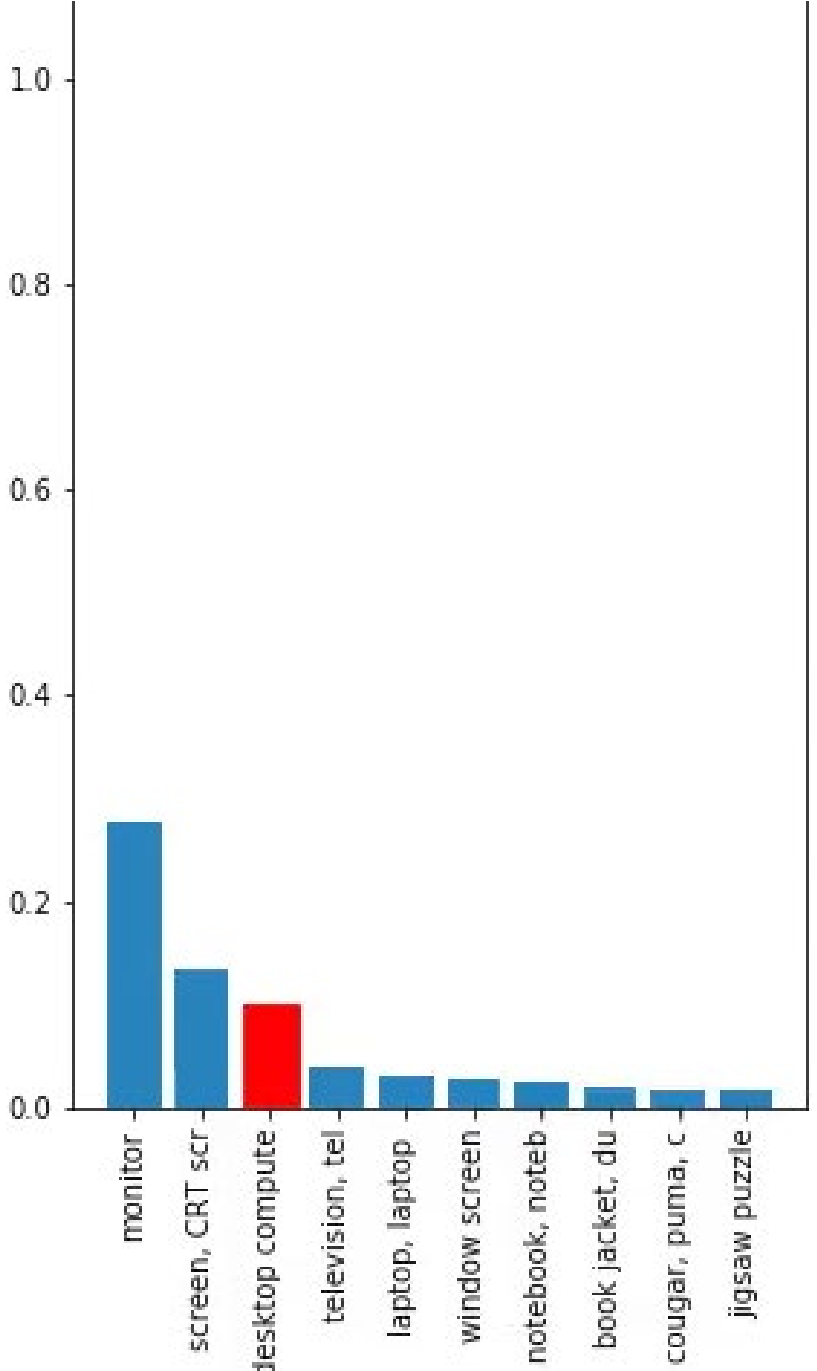
Szegedy et al. – *Intriguing Properties of Neural Networks,* ICLR '14 (https://arxiv.org/abs/1312.6199v4)

**Which class is this?**

**Cat**

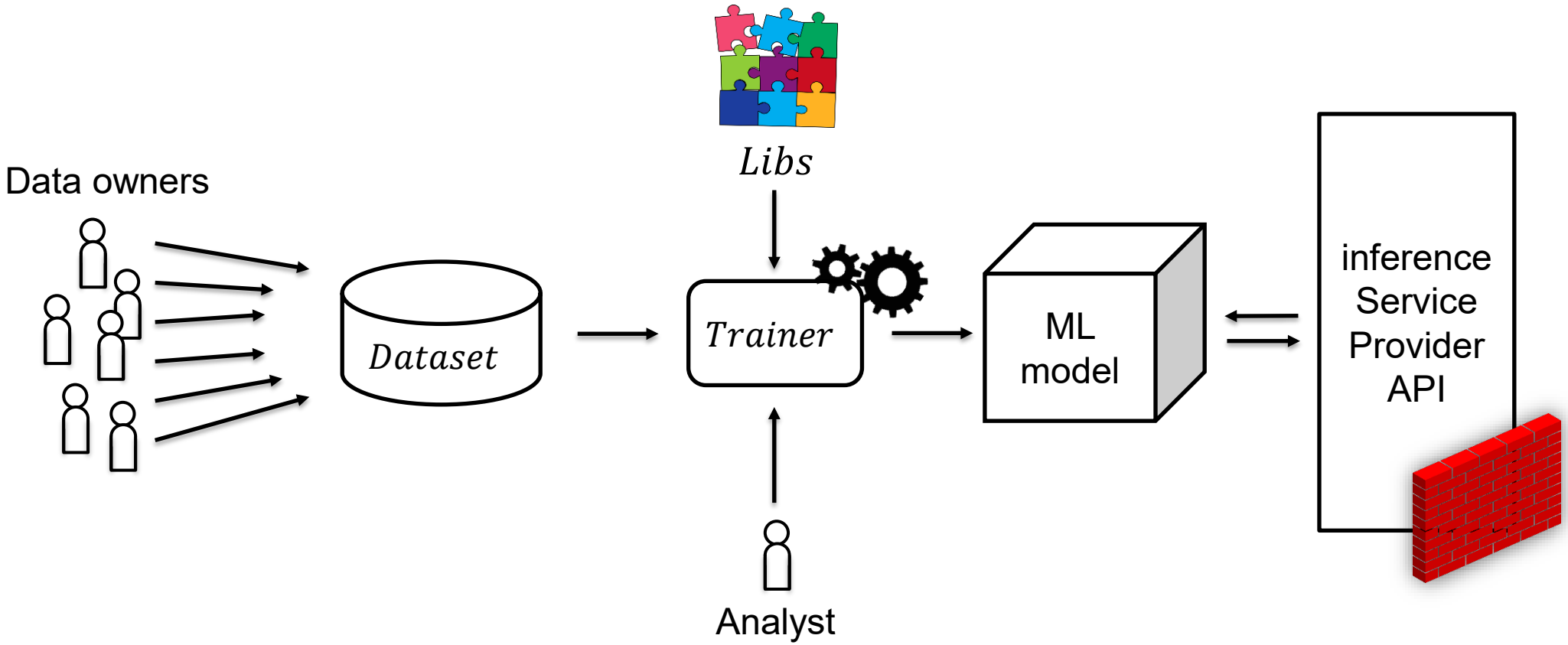
**Which class is this?**

**Desktop computer**



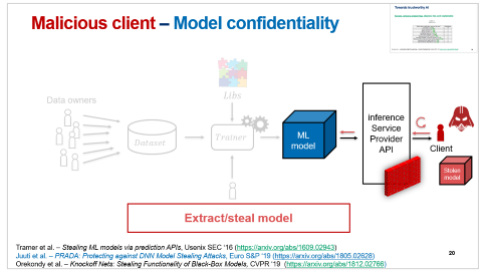

Athalye et al. – *Synthesizing Robust Adversarial Examples*, ICML '2019 (https://blog.openai.com/robust-adversarial-inputs/)

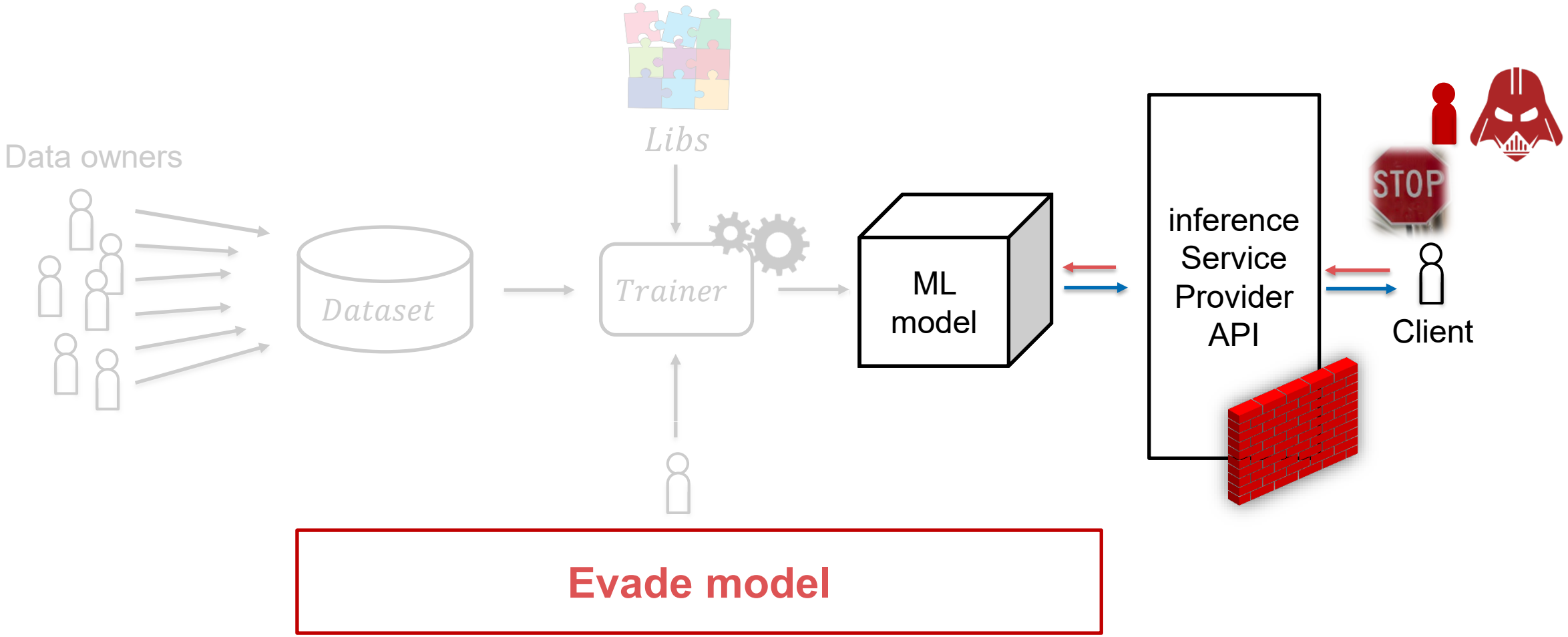
# Machine Learning pipeline




**Where is the adversary? What is its target?**

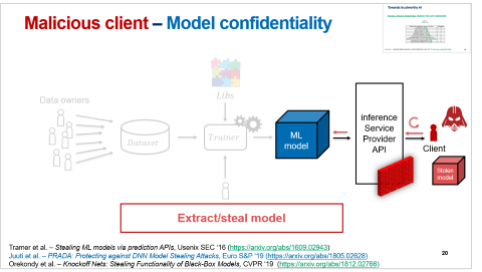
# Compromised input – Model integrity

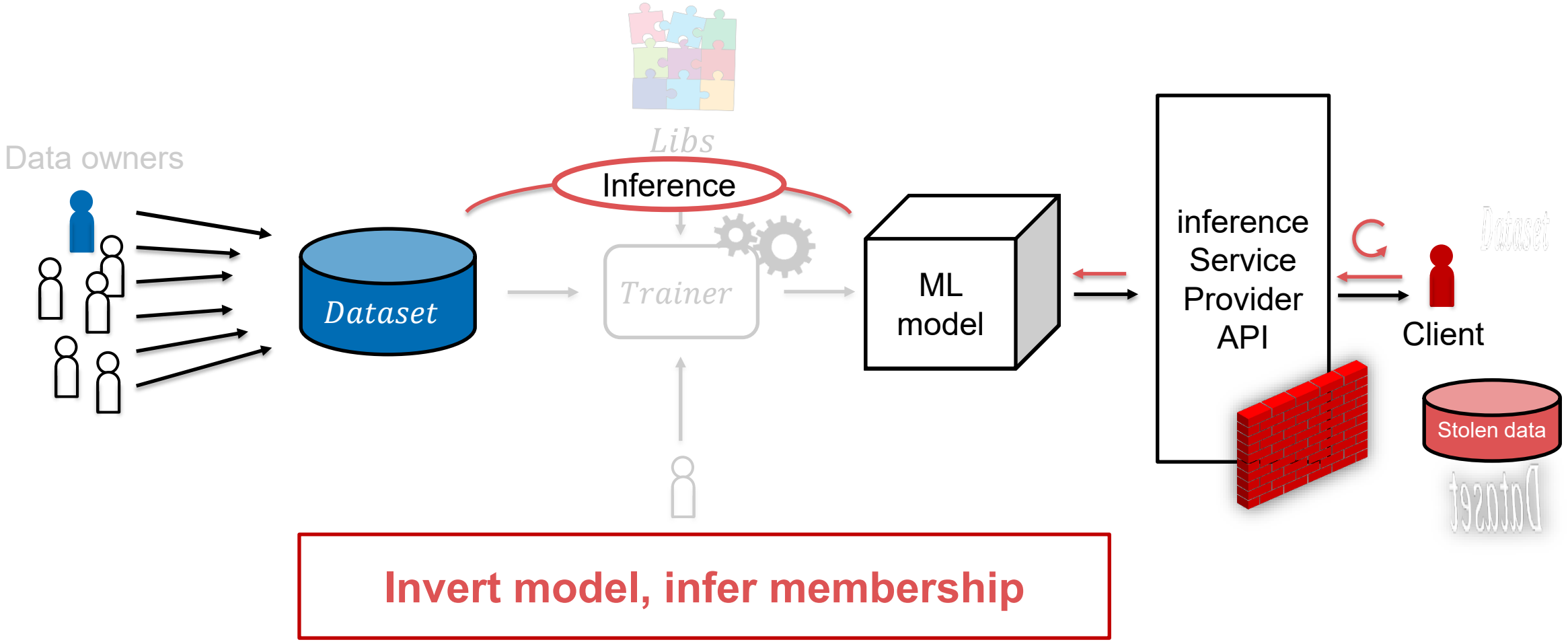



Szegedy et al. – *Intriguing Properties of Neural Networks,* ICLR '14 (https://arxiv.org/abs/1312.6199v4)
Dalvi et al. – *Adversarial Classification,* KDD '04 (https://dl.acm.org/doi/10.1145/1014052.1014066)

# Malicious client – Training data privacy

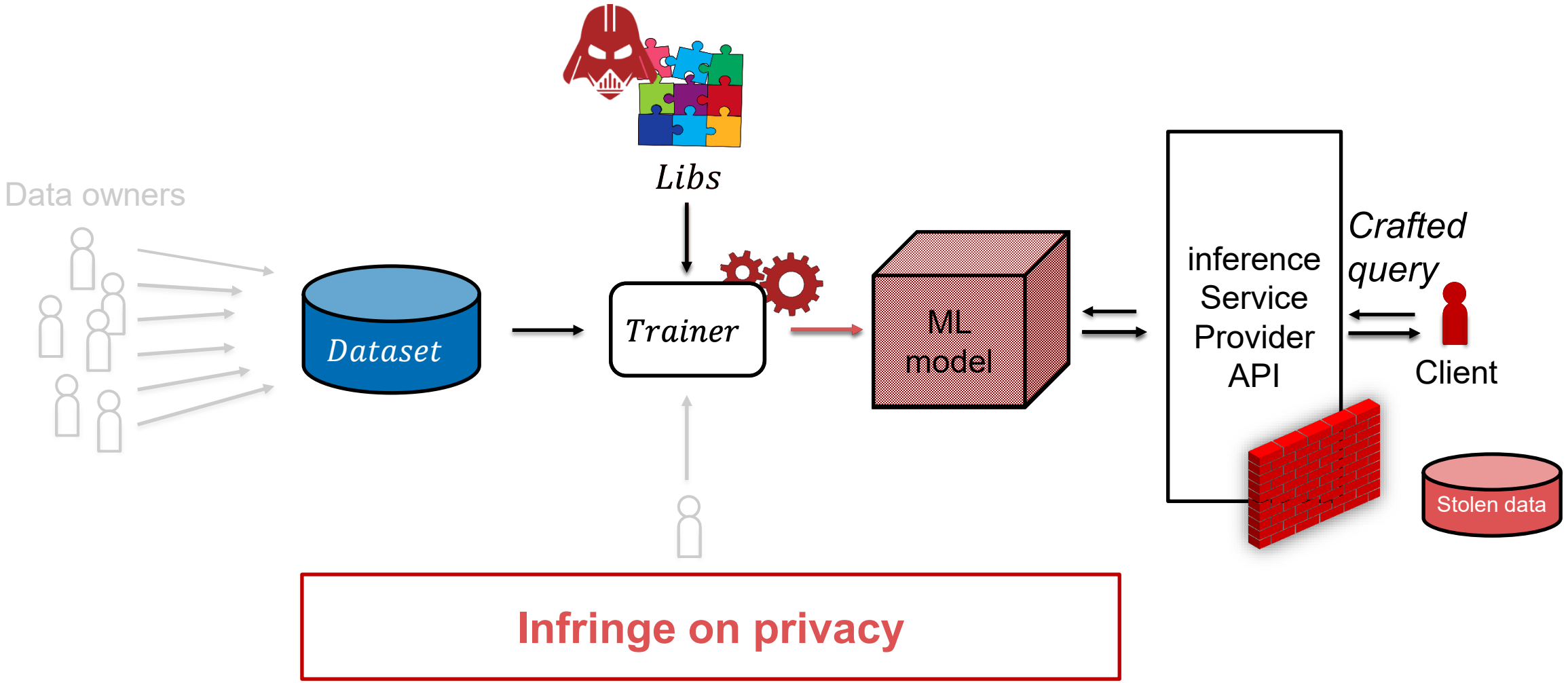



**Invert model, infer membership**

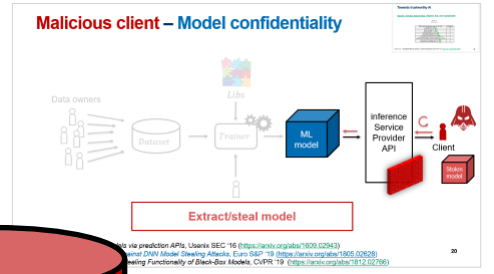
Shokri et al. – *Membership Inference Attacks Against Machine Learning Models,* IEEE S&P '16 (https://arxiv.org/pdf/1610.05820.pdf)
Fredrikson et al. – *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, ACM CCS '15 (https://doi.org/10.1145/2810103.2813677)

# Compromised toolchain – Training data privacy

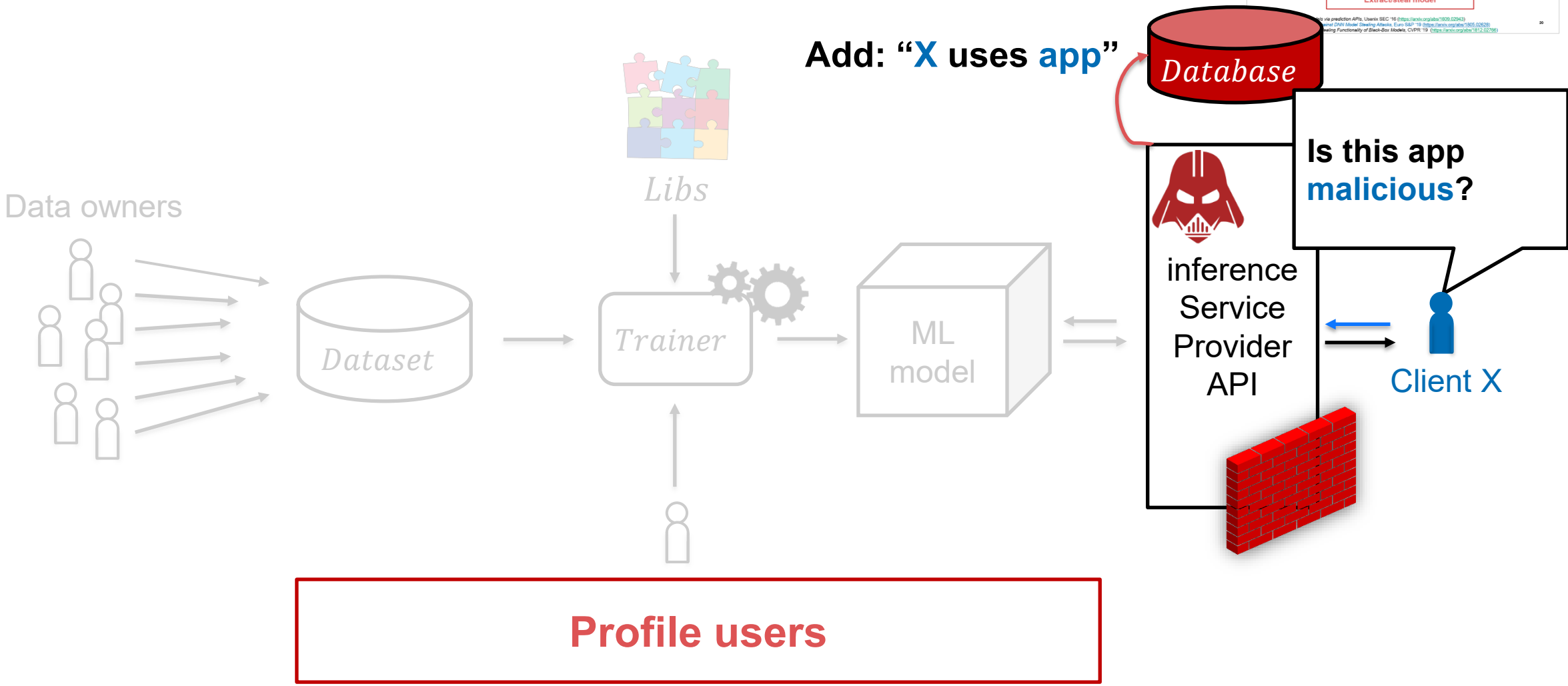



Song et al. – *Machine Learning models that remember too much*, ACM CCS '17 (https://arxiv.org/abs/1709.07886)
Hitja et al. – *Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning*, ACM CCS '17 (http://arxiv.org/abs/1702.07464)

# Malicious inference service – User profiles

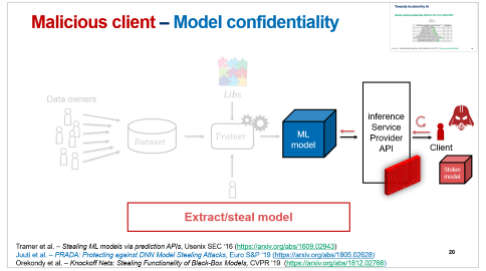


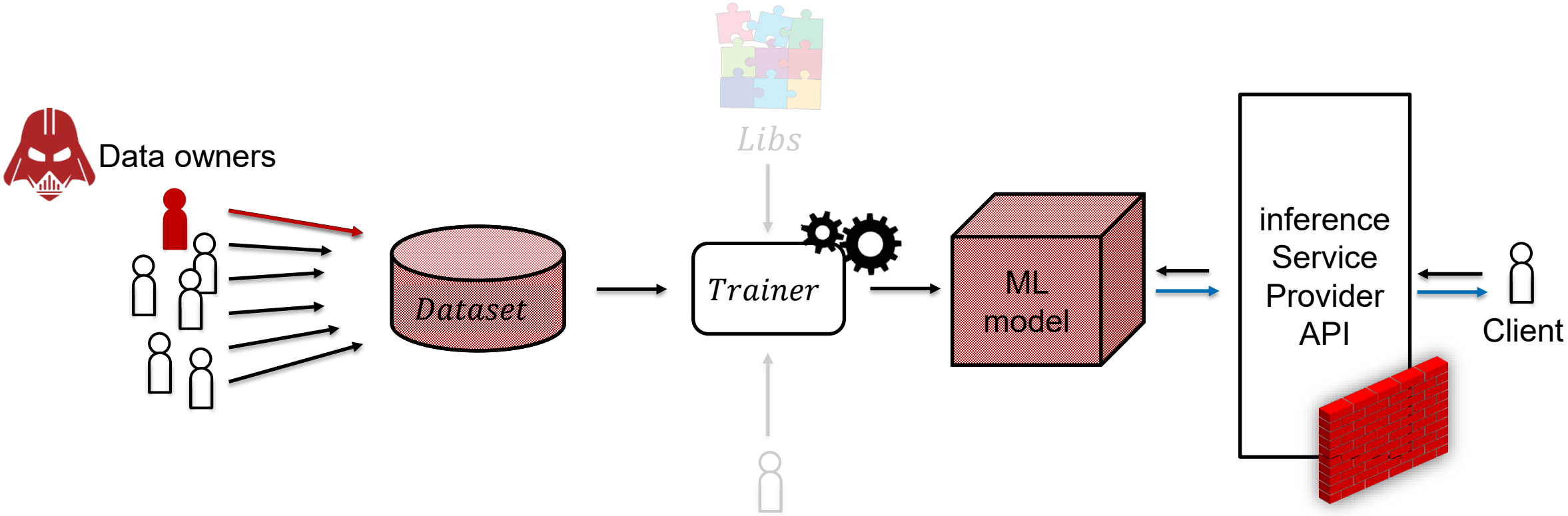

Add: "X uses app"



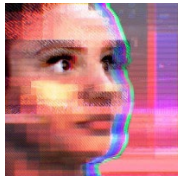

Malmi and Weber – *You are what apps you use Demographic prediction based on user's apps*, ICWSM '16 (https://arxiv.org/abs/1603.00059)
Dowlin et al. – *CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy*, ICML '16 (https://dl.acm.org/doi/10.5555/3045390.3045413 )
Liu et al. – *Oblivious Neural Network Predictions via MiniONN Transformations*, ACM CCS '17 (https://ssg.aalto.fi/research/projects/mlsec/ppml/)

# Malicious data owner – Model integrity



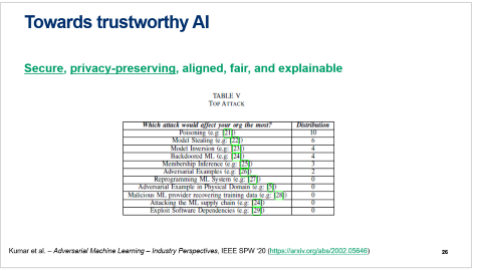



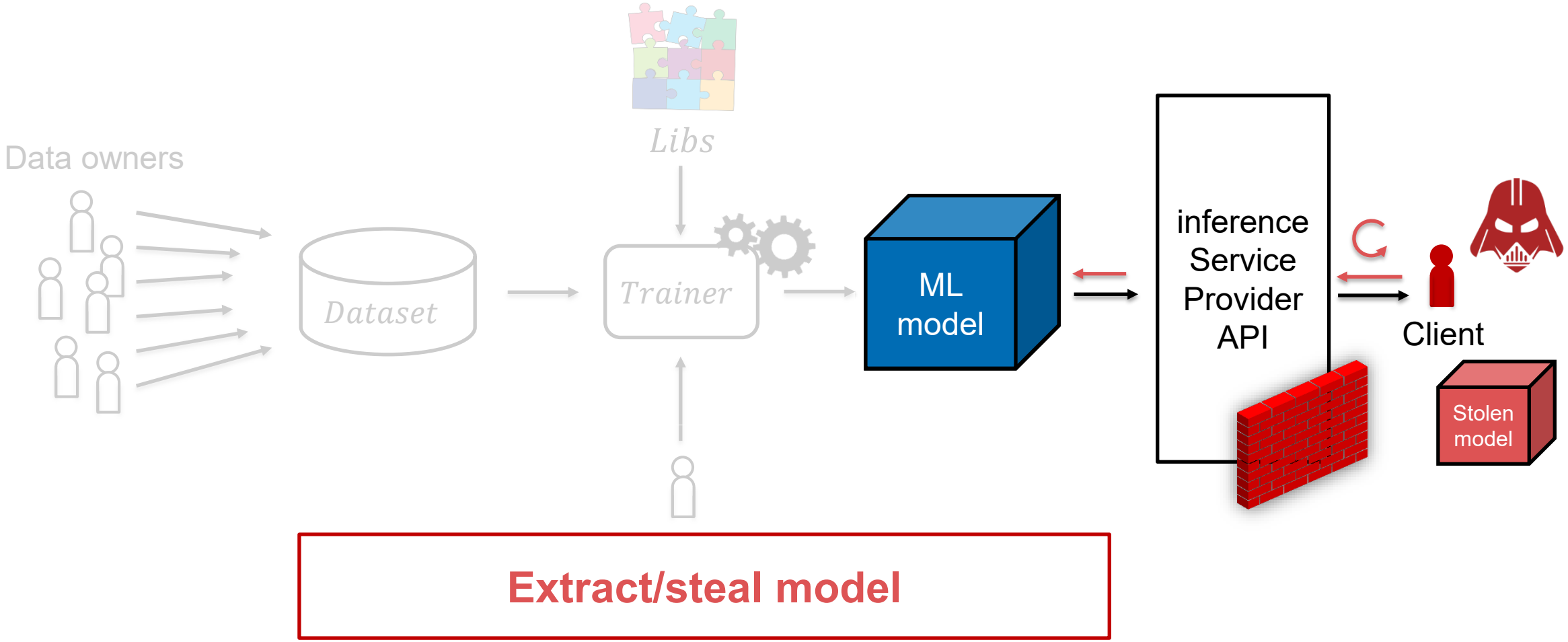

**Influence ML model (model poisoning)**



https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot
https://www.theguardian.com/technology/2017/nov/07/youtube-accused-violence-against-young-children-kids-content-google-pre-school-abuse

# Malicious client – Model confidentiality




**Extract/steal model**

Tramer et al. – *Stealing ML models via prediction APIs*, Usenix SEC '16 (https://arxiv.org/abs/1609.02943)
Juuti et al. – *PRADA: Protecting against DNN Model Stealing Attacks*, Euro S&P '19 (https://arxiv.org/abs/1805.02628)
Orekondy et al. – *Knockoff Nets: Stealing Functionality of Black-Box Models,* CVPR '19 (https://arxiv.org/abs/1812.02766)

# Towards trustworthy AI

## Secure, privacy-preserving, …

TABLE V
TOP ATTACK

| *Which attack would affect your org the most?* | *Distribution* |
|---|---|
| Poisoning (e.g: [21]) | 10 |
| Model Stealing (e.g: [22]) | 6 |
| Model Inversion (e.g: [23]) | 4 |
| Backdoored ML (e.g: [24]) | 4 |
| Membership Inference (e.g: [25]) | 3 |
| Adversarial Examples (e.g: [26]) | 2 |
| Reprogramming ML System (e.g: [27]) | 0 |
| Adversarial Example in Physical Domain (e.g: [5]) | 0 |
| Malicious ML provider recovering training data (e.g: [28]) | 0 |
| Attacking the ML supply chain (e.g: [24]) | 0 |
| Exploit Software Dependencies (e.g: [29]) | 0 |

Kumar et al. – *Adversarial Machine Learning – Industry Perspectives*, IEEE SPW '20 (https://arxiv.org/abs/2002.05646)

# Unintended interactions between defenses and risks

**Prior work explored defenses to mitigate specific risks**

- Defenses typically evaluated only vs. those specific risks they protect against

**But practitioners need to deploy multiple defenses simultaneously**

- Can two defenses interact negatively with each other?
- Does a defense exacerbate or ameliorate some other (unrelated) risk?

Blog article: https://blog.ssg.aalto.fi/2024/05/unintended-interactions-among-ml.html

Szyller and Asokan – *Conflicting Interactions Among Protections Mechanisms for Machine Learning Models*, AAAI '23 (https://arxiv.org/abs/2207.01991)
Duddu, Szyller, and Asokan - *SoK: Unintended Interactions among Machine Learning Defenses and Risks,* IEEE S&P '24. https://arxiv.org/abs/2312.04542

# Is malicious adversarial behaviour the only concern?


BBC Sign in Home News Sport Reel Worklife

NEWS

Home | US Election | Coronavirus | Video | World | UK | Business | Tech | Science | Stories | Entertainment

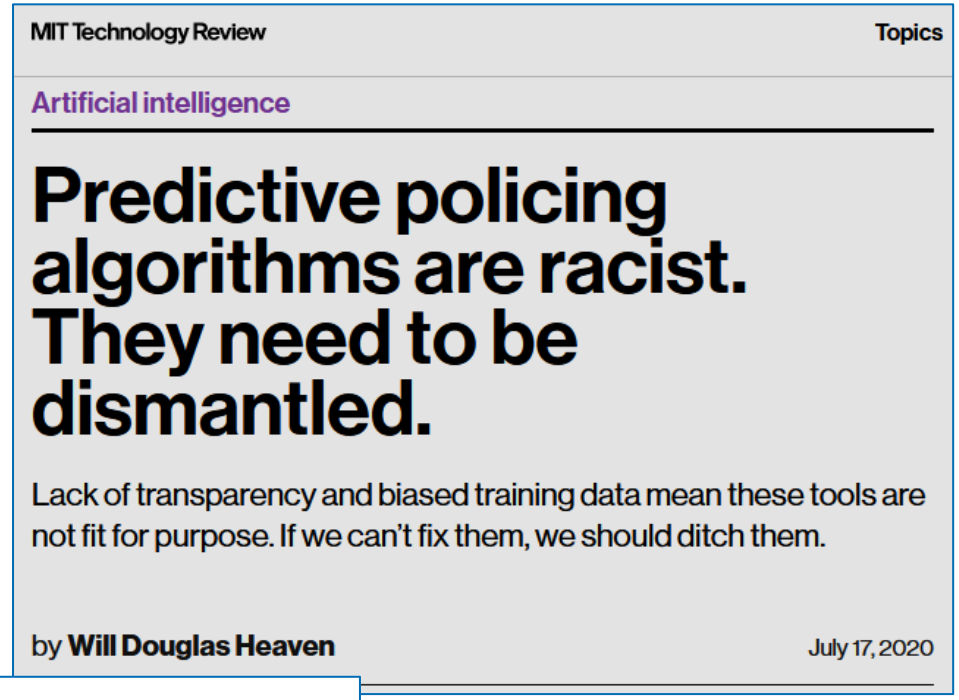
Tech

Twitter investigates racial bias in image previews

19 hours ago

One user found that Twitter seemed to favour showing Mitch McConnell's face over Barack Obama's

https://www.bbc.com/news/technology-54234822?fbclid=IwAR1T41_HR6lIuMKGRJbJdDrdpKdyAi5mhQSdzs0QLDso41T-SR3wJfs


MIT Technology Review Topics

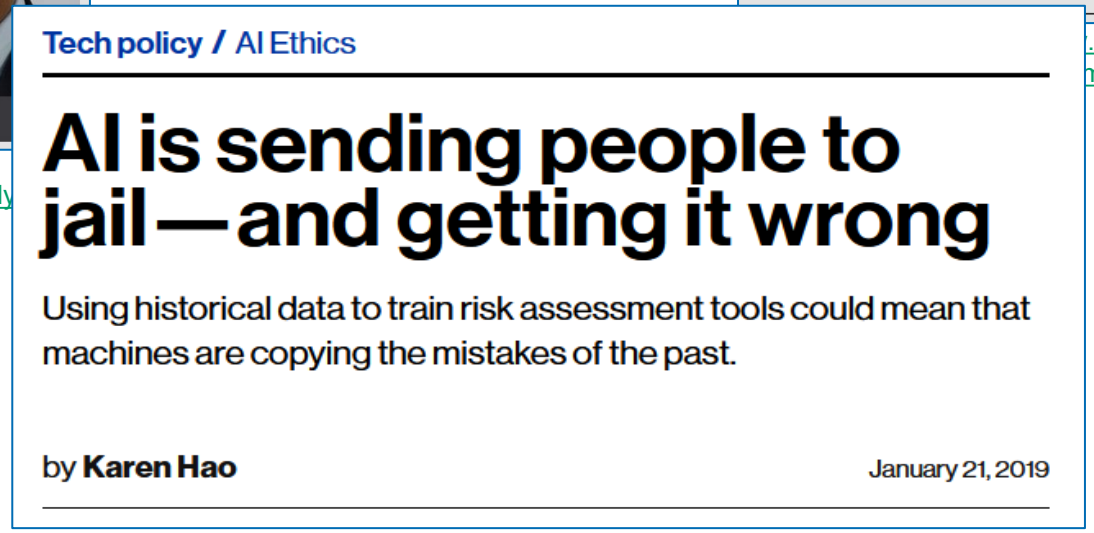
Artificial intelligence

Predictive policing algorithms are racist. They need to be dismantled.

Lack of transparency and biased training data mean these tools are not fit for purpose. If we can't fix them, we should ditch them.

by Will Douglas Heaven July 17, 2020

.com/2020/07/17/1005396/predictive-policing-machine-learning-bias-criminal-justice/


Tech policy / AI Ethics

AI is sending people to jail—and getting it wrong

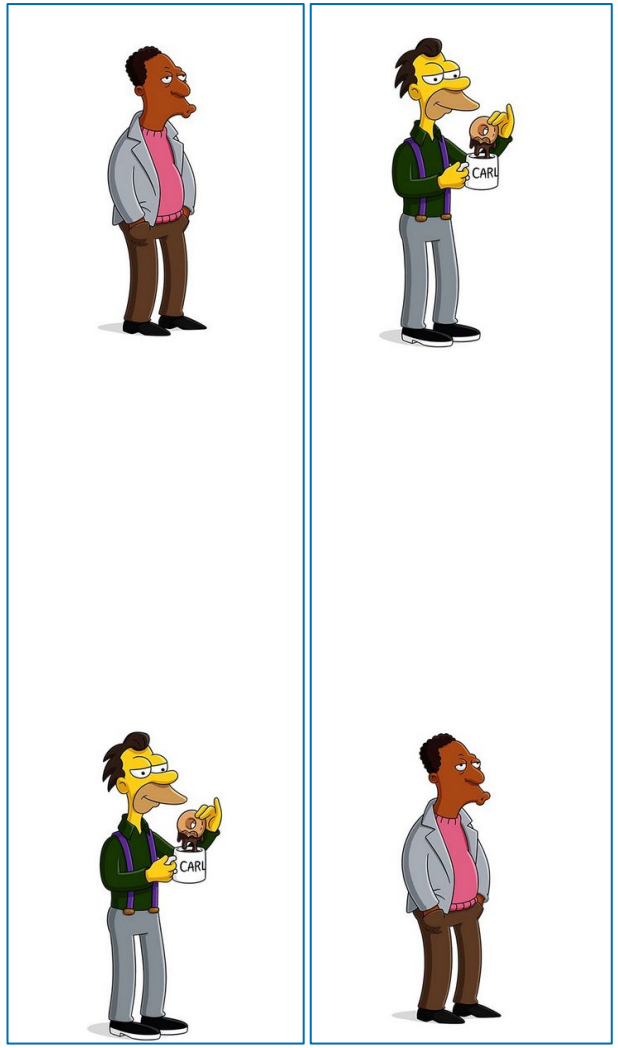
Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by Karen Hao January 21, 2019

https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai /

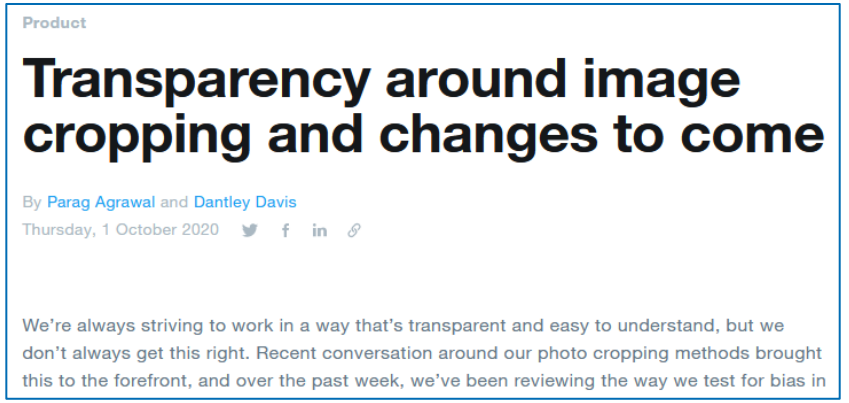
# Measures of accuracy are flawed, too


https://twitter.com/_jsimonovski/status/1307542747197239296





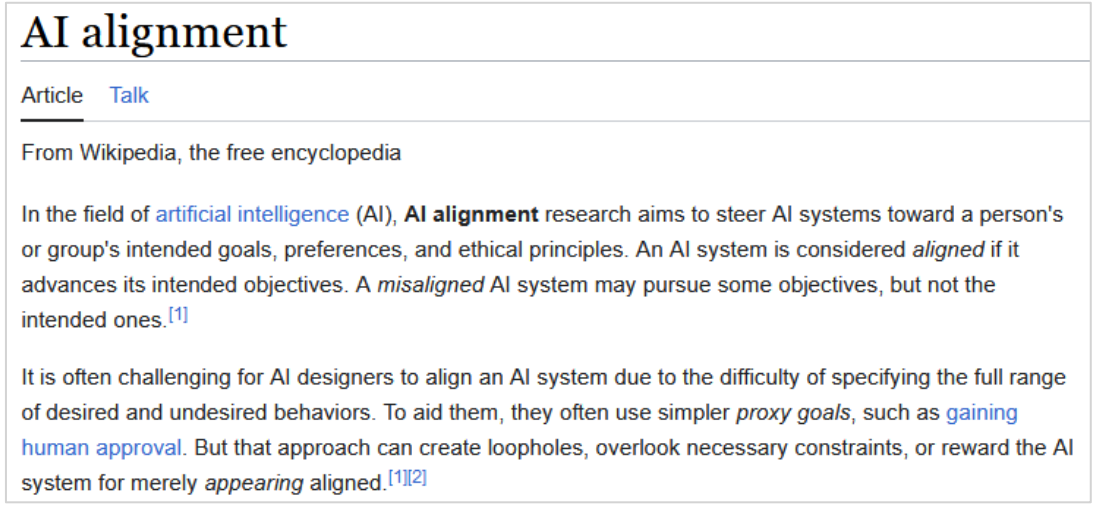
https://twitter.com/TwitterComms/status/1307739940424359936




https://blog.twitter.com/official/en_us/topics/product/2020/transparency-image-cropping.html

# Other AI trustworthiness concerns

## Unaligned AI



AI alignment

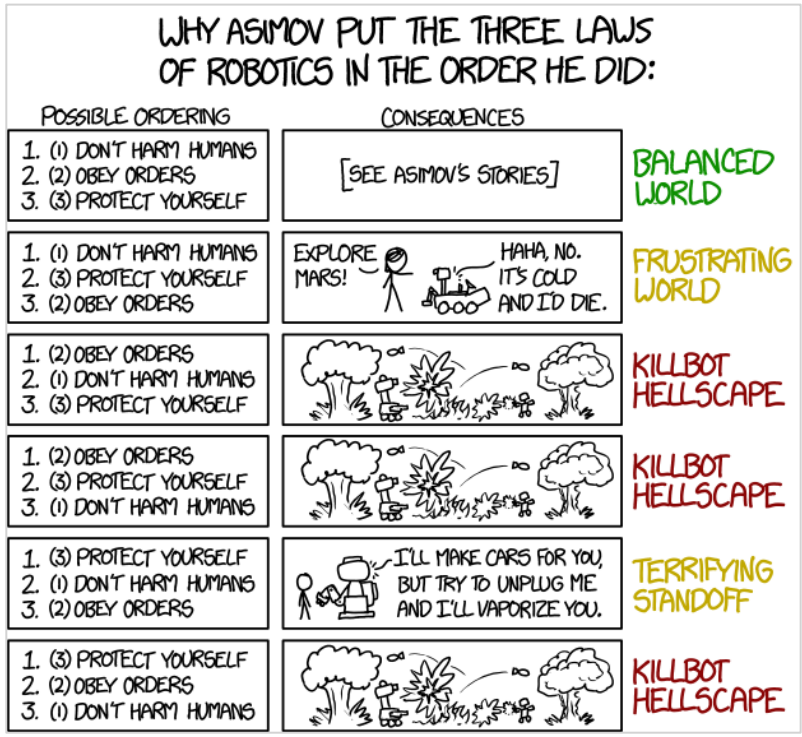
Article Talk

From Wikipedia, the free encyclopedia

In the field of artificial intelligence (AI), **AI alignment** research aims to steer AI systems toward a person's or group's intended goals, preferences, and ethical principles. An AI system is considered *aligned* if it advances its intended objectives. A *misaligned* AI system may pursue some objectives, but not the intended ones.[1]

It is often challenging for AI designers to align an AI system due to the difficulty of specifying the full range of desired and undesired behaviors. To aid them, they often use simpler *proxy goals*, such as gaining human approval. But that approach can create loopholes, overlook necessary constraints, or reward the AI system for merely *appearing* aligned.[1][2]

https://en.wikipedia.org/wiki/AI_alignment




https://xkcd.com/1613/

## AI-enabled fraud

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM › PRESIDENTIAL ACTIONS

https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

# Takeaways

**Trustworthy AI-based systems must address security & privacy**

*Active research topic*

**Other related concerns: fairness, explainability, alignment, …**

**AI-enabled fraud is a growing concern**

**Our research topics**

**ML security/privacy:**

**ML ownership resolution, Conflicting ML defenses, ML property attestation, robust concept removal in gen AI**

**Platform security:**

**hardware-assisted run-time security, secure outsourced computing**

**Open (postdoc, grad student) positions to help lead our work: ML security/privacy, platform security**
**https://asokan.org/asokan/research/SecureSystems-open-positions-Jan2024.php**