UNIVERSITY OF WATERLOO

Confidence in AI

Can we trust AI-based systems?

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

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

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 design/use hardware assistance to secure software?

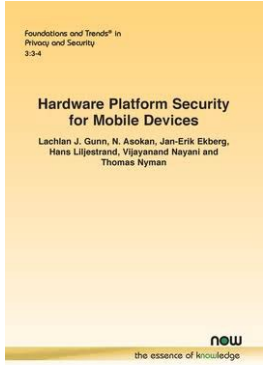
https://ssg-research.github.io/

Platform security research

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

Hardware assisted trusted execution environments (TEEs)

CCS 2019 keynote[1] https://youtu.be/hHYoGn5PSl4



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

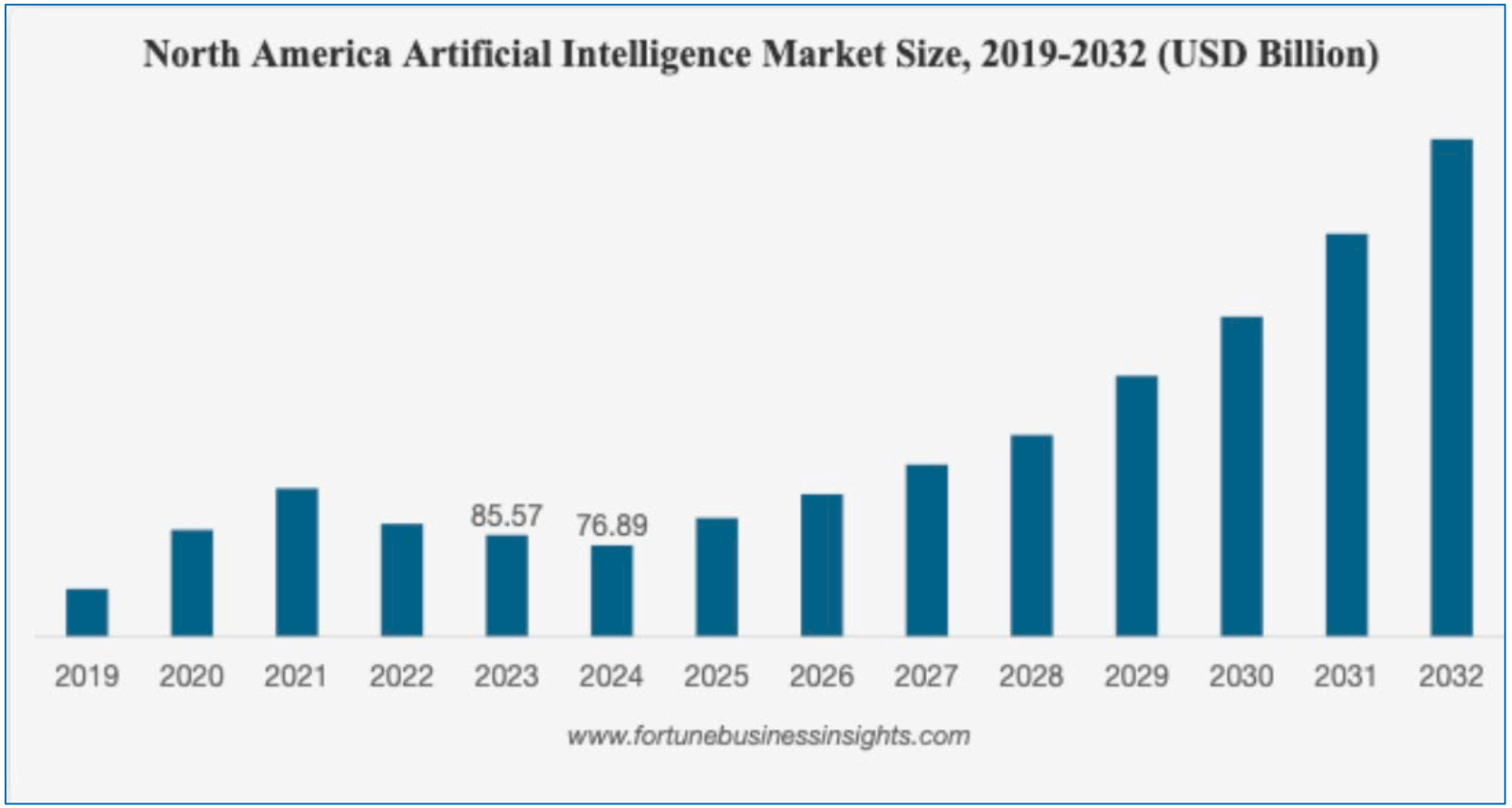
Novel hardware security mechanisms

- 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

VICE NEWSLETTERS

Tech

Dozens of Cities Have Secretly Experimented With Predictive Policing Software

By Caroline Haskins **February 6, 2019, 11:00am**

https://www.vice.com/en_us
experimented-with-predictiv

Forbes Sign In

How AI Is Uprooting Recruiting

By Falon Fatemi , Contributor. Forbes Contributor covering the future of...

Follow Author

Oct 31, 2019, 02:42pm EDT

https://www.forbes.com/sites/falonfatemi/2019/10/31/how-ai-is-uprooting-recruiting/

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 nee 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-t

Challenges in making AI trustworthy

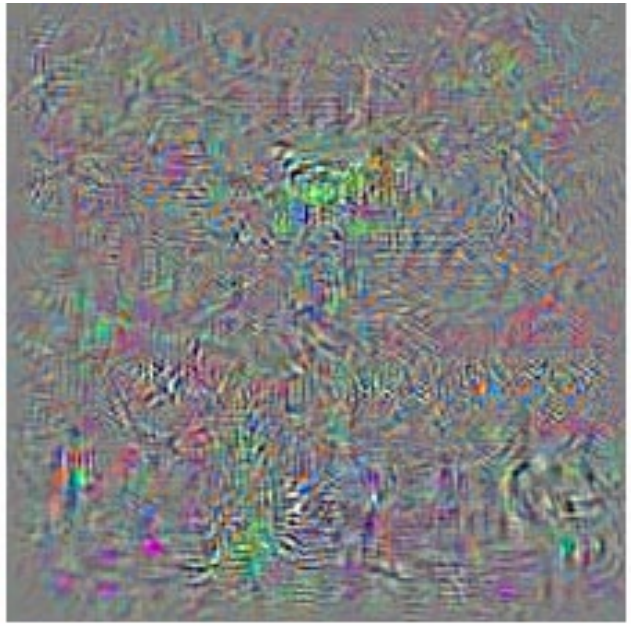
Security concerns

Privacy concerns

[Other concerns: fairness, explainability, alignment]

Evading machine learning models

+ 0.1·



=

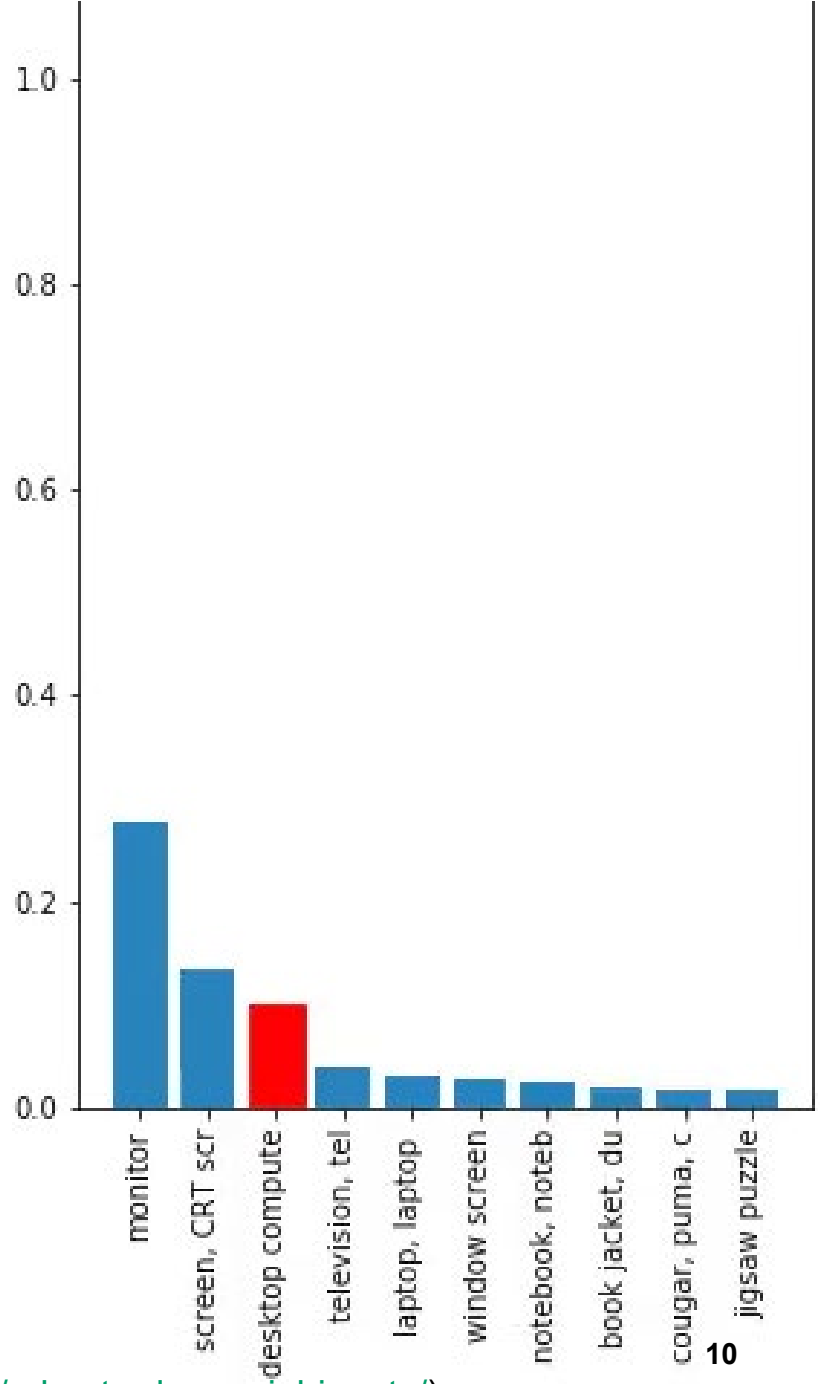
Which class is this?
School bus

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/)

智能系统安全实验室
UBIQUITOUS SYSTEM SECURITY LAB.

DolphinAttack: Inaudible Voice command

Guoming Zhang Chen Yan Xiaoyu Ji

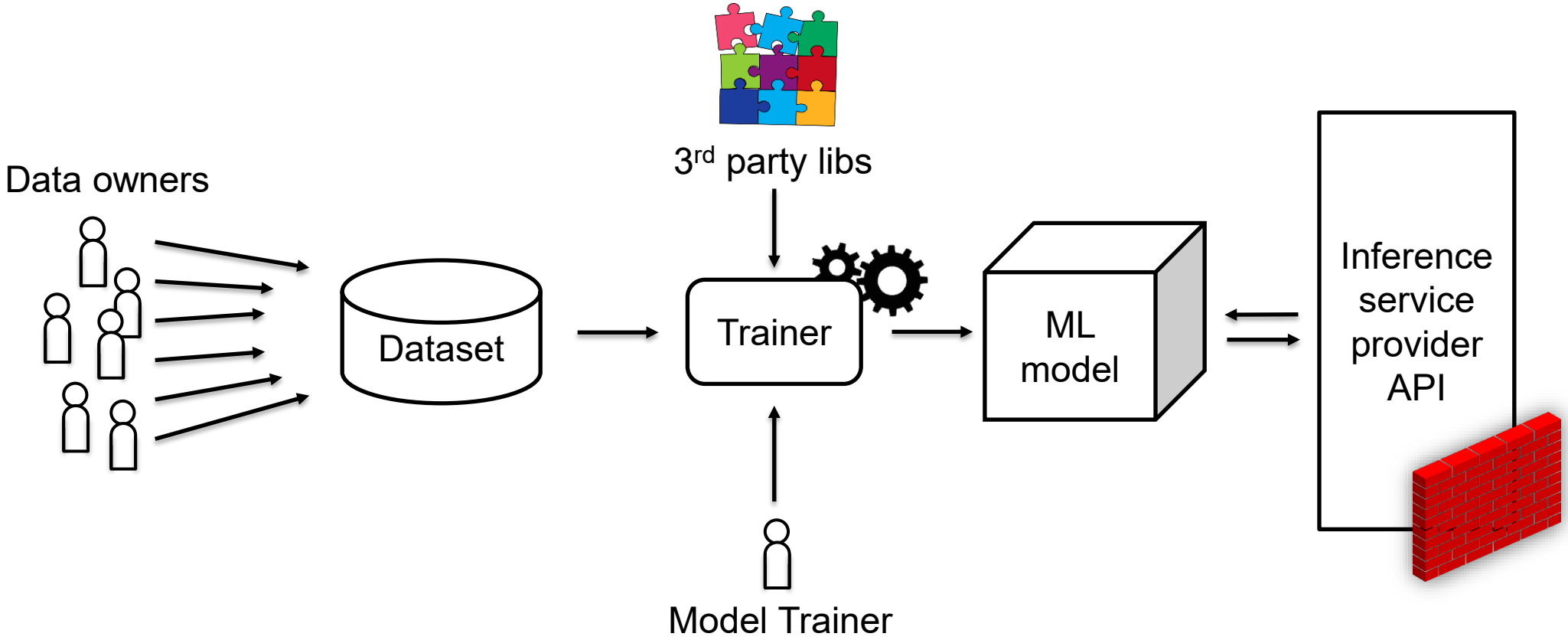
Tianchen Zhang Taimin Zhang Wenyuan Xu

Zhejiang University

ACM CCS 2017

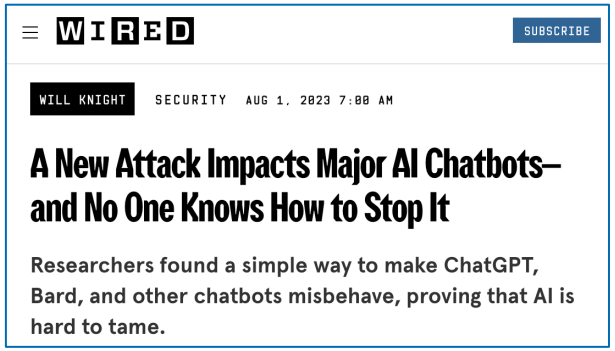
Zhang et al. – *DolphinAttack: Inaudible Voice Commands*, ACM CCS ‘17 (https://arxiv.org/abs/1708.09537)

Machine Learning pipeline



Where is the adversary? What is its target?

Compromised input – Model integrity

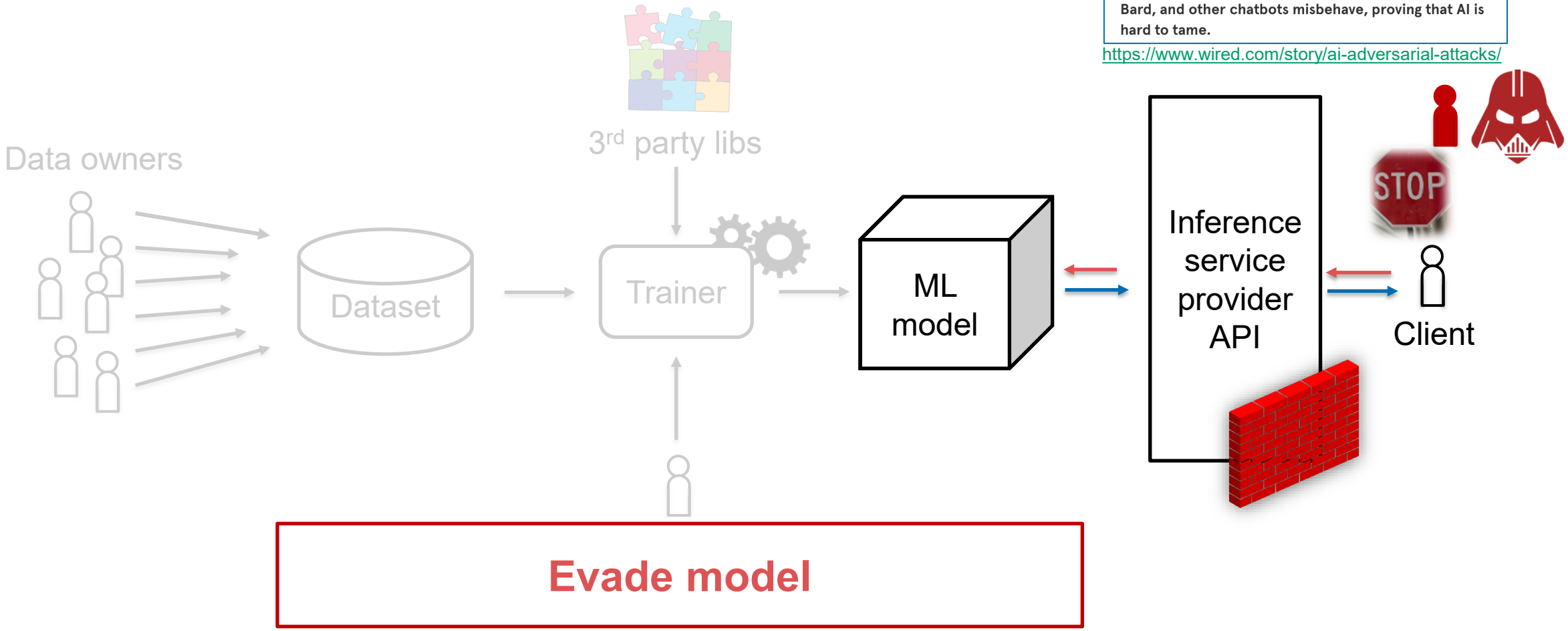

WIRED SUBSCRIBE

WILL KNIGHT SECURITY AUG 1, 2023 7:00 AM

A New Attack Impacts Major AI Chatbots—and No One Knows How to Stop It

Researchers found a simple way to make ChatGPT, Bard, and other chatbots misbehave, proving that AI is hard to tame.

https://www.wired.com/story/ai-adversarial-attacks/



Madry et al. – *Towards Deep Learning Models Resistant to Adversarial Attacks,* ICLR '18 (https://arxiv.org/abs/1706.06083)
Carlini & Wagner. – *Towards Evaluating the Robustness of Neural Networks,* IEEE S&P '17 (https://arxiv.org/abs/1608.04644)

Malicious client – Training data privacy

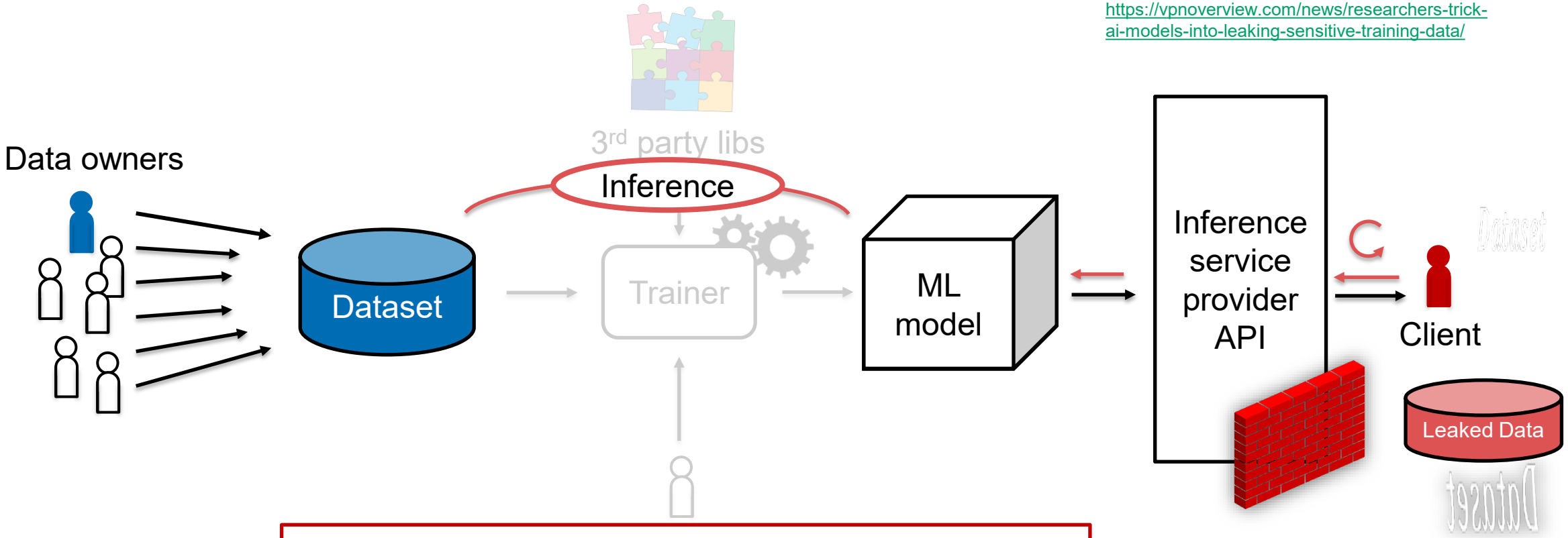

vpnoverview

Researchers Trick AI Models Into Leaking Sensitive Training Data

Published: 12-01-2023

Mirza Silajdzic
Senior News Journalist

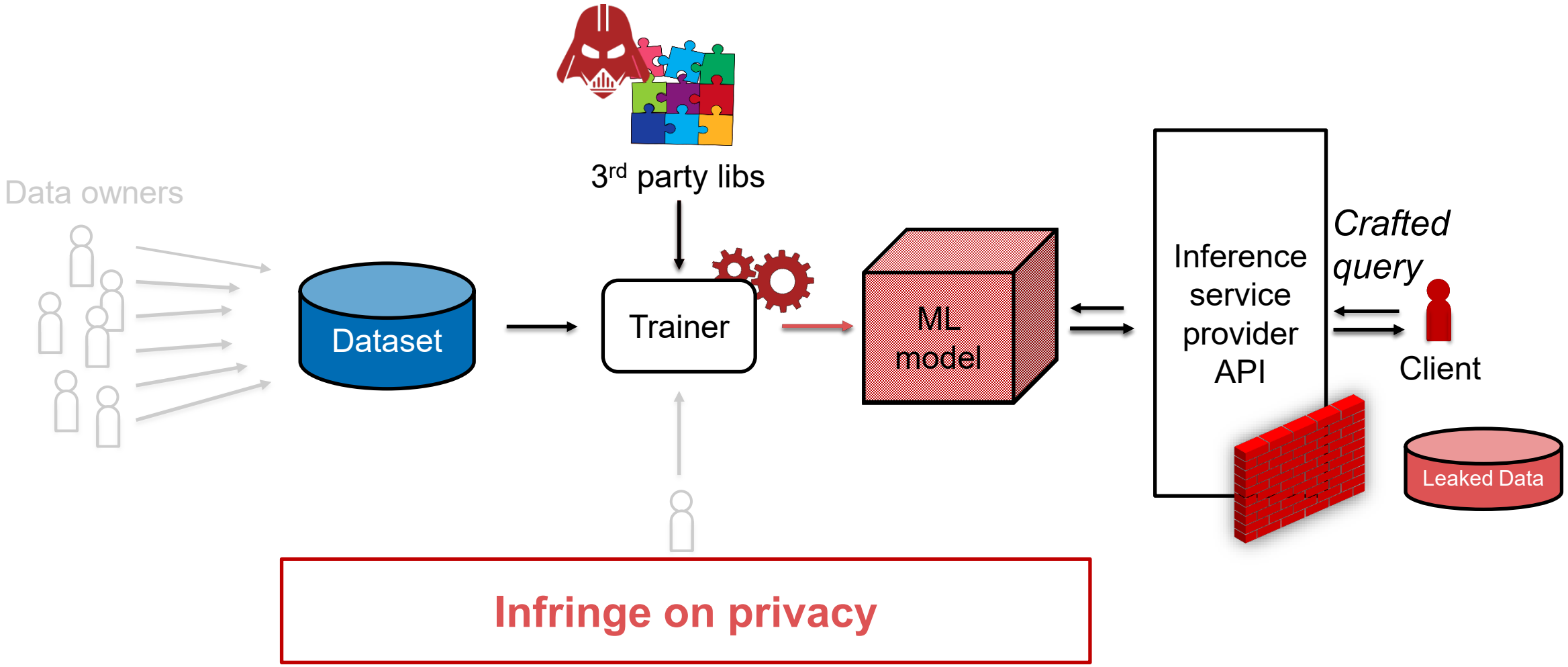
https://vpnoverview.com/news/researchers-trick-ai-models-into-leaking-sensitive-training-data/



Invert model, infer membership

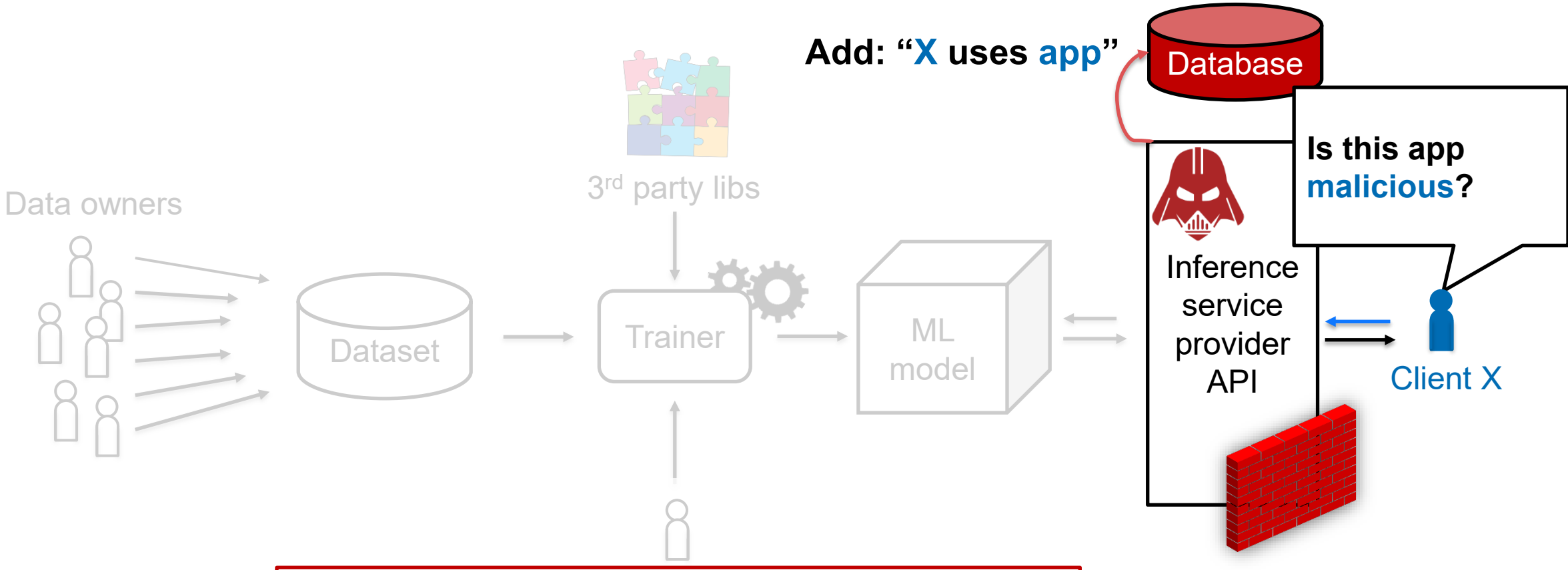
Carlini et al. – *Membership Inference Attacks From First Principles*, IEEE S&P '22 (https://arxiv.org/abs/2112.03570)
Jayaram & Evans – *Are Attribute Inference Attacks Just Imputation?*, ACM CCS '22 (https://arxiv.org/abs/2209.01292)
Carlini et al. – *Extracting Training Data from Large Language Models*, USENIX SEC '21 (https://arxiv.org/abs/2012.07805)
Suri et al. – *Dissecting Distribution Inference*, SaTML '23 (https://arxiv.org/abs/2212.07591)

Compromised toolchain – Training data privacy



Song et al. – *Machine Learning models that remember too much*, ACM CCS '17 (https://arxiv.org/abs/1709.07886)
Bagdasararyan & Shmatikov – *Blind Backdoors in Deep Learning Models*, USENIX SEC '21 (https://arxiv.org/abs/2005.03823)

Malicious inference service – Private inference

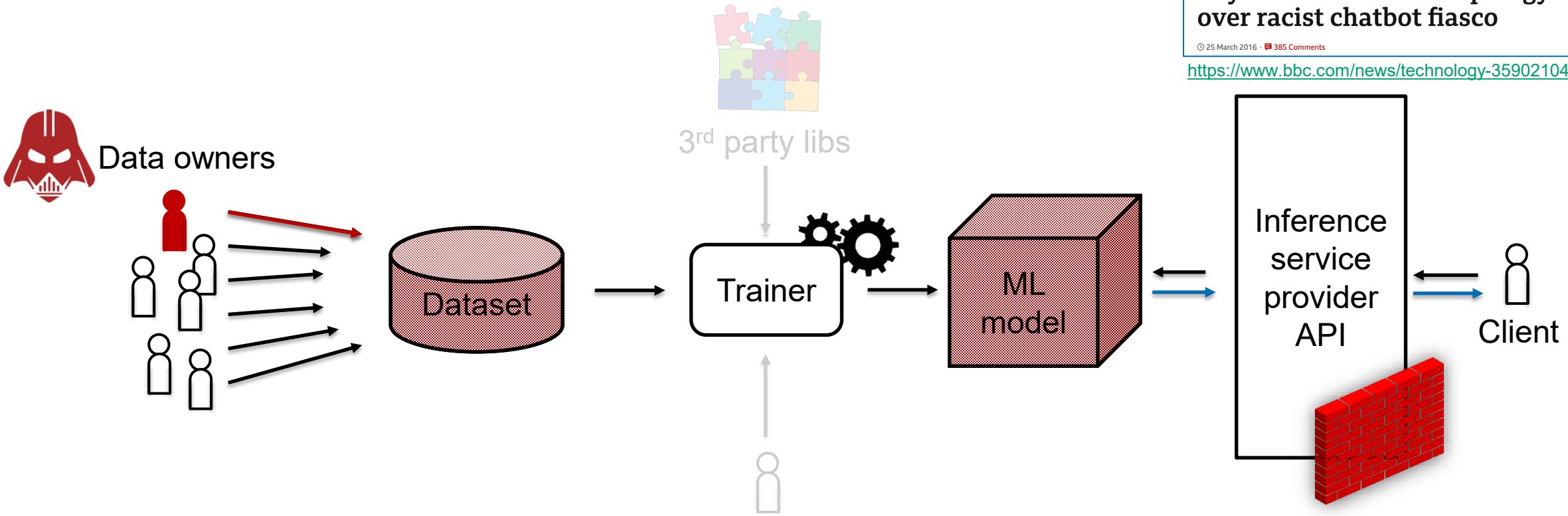


Misuse client input/ouput

Malmi and Weber – *You are what apps you use Demographic prediction based on user's apps*, ICWSM '16 (https://arxiv.org/abs/1603.00059)
Liu et al. – *Oblivious Neural Network Predictions via MiniONN Transformations*, ACM CCS '17 (https://ssg.aalto.fi/research/projects/mlsec/ppml/)
Zhang et al. – *Secure Transformer Inference Made Non-interactive*, NDSS '25 (https://www.ndss-symposium.org/wp-content/uploads/2025-868-paper.pdf)

Malicious data owner – Model integrity

https://www.bbc.com/news/technology-35902104



Influence ML model (model poisoning)

Gu et al. – *BadNets: Evaluating Backdooring Attacks on Deep Neural Networks*, IEEE Access '19 (https://ieeexplore.ieee.org/document/8685687)
Li et al. – *Anti-Backdoor Learning: Training Clean Models on Poisoned Data*, NeurIPS '21 (https://arxiv.org/abs/2110.11571)

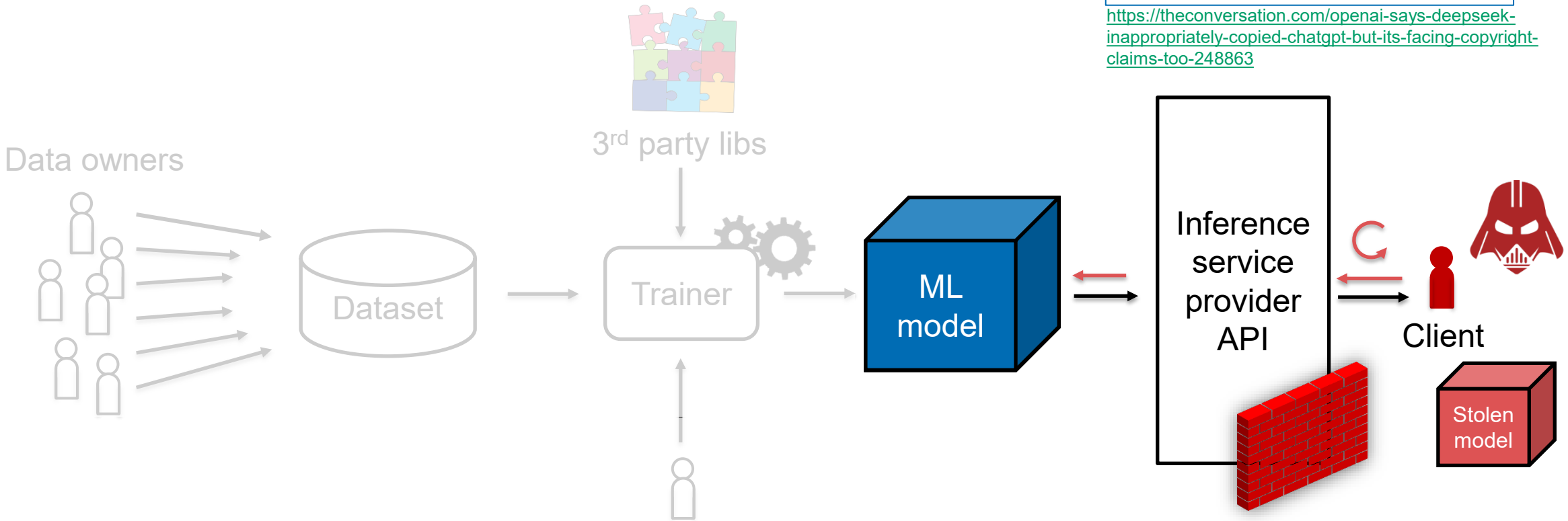
Malicious client – Model confidentiality

OpenAI says DeepSeek 'inappropriately' copied ChatGPT – but it's facing copyright claims too

Published: February 4, 2025 2.10pm EST

Lea Frermann, Shaanan Cohney, *The University of Melbourne*

https://theconversation.com/openai-says-deepseek-inappropriately-copied-chatgpt-but-its-facing-copyright-claims-too-248863



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)
Carlini et al. – *Stealing part of a production language model,* ICML '24 (https://arxiv.org/abs/2403.06634)

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?[1]
- Does a defense exacerbate or ameliorate some other (unrelated) risk?[2] **Distinguished Paper Award**

Conjecture: overfitting and memorization are influence defenses and risks[2][3]

- Effective defenses may induce, reduce or rely on overfitting or memorization
- Risks tend to exploit overfitting or memorization
- Underlying factors that influence memorization/overfitting can be identified

Recently built a toolkit, Amulet, for comparative evaluation of attacks & defenses[4]

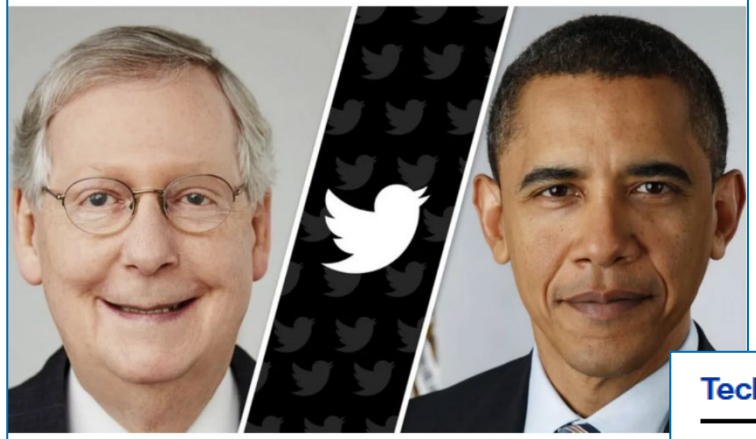
[1] Szyller and Asokan – *Conflicting Interactions Among Protections Mechanisms for Machine Learning Models*, AAAI '23 (https://arxiv.org/abs/2207.01991)
[2] Duddu, Szyller, and Asokan - *SoK: Unintended Interactions among Machine Learning Defenses and Risks,* IEEE S&P '24 (https://arxiv.org/abs/2312.04542)
[3] Blog article: https://blog.ssg.aalto.fi/2024/05/unintended-interactions-among-ml.html
[4] Amulet repo: https://github.com/ssg-research/amulet

Is malicious adversarial behaviour the only concern?

BBC

Twitter investigates racial bias in image previews

21 September 2020



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

https://www.bbc.com/news/technology-54234822?fbclid=IwAR1T41_HR6IluMKGRJbJdDrdpKdyAi5mhQSdzs0QLDso41T-SR3wJfs

MIT Technology Review SUBSCRIBE

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

https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-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.

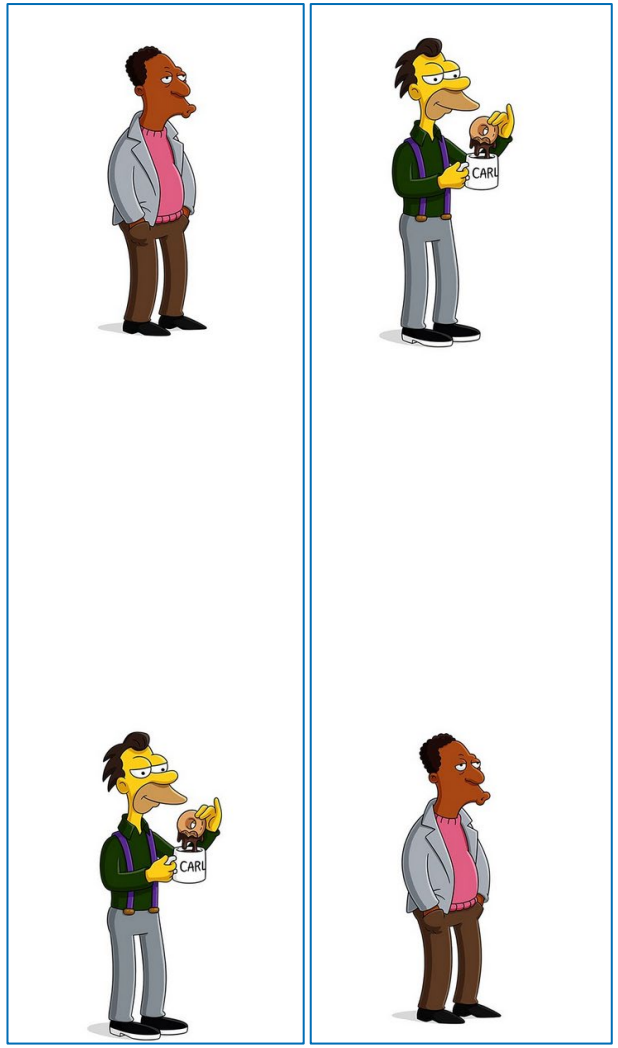
by Karen Hao

January 21, 2019

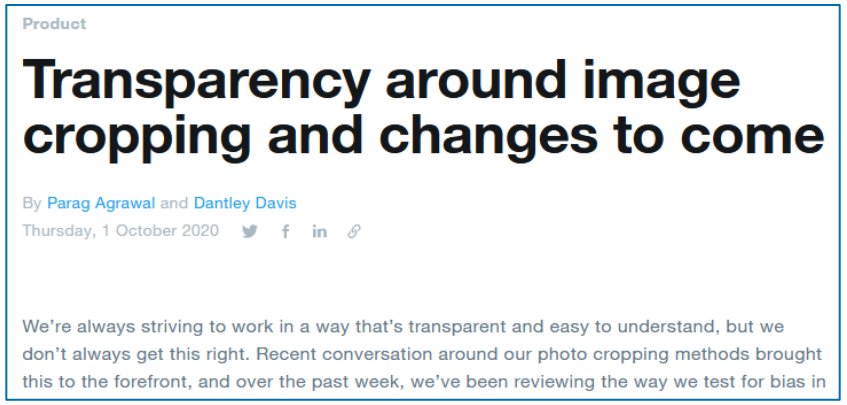
https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai /

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

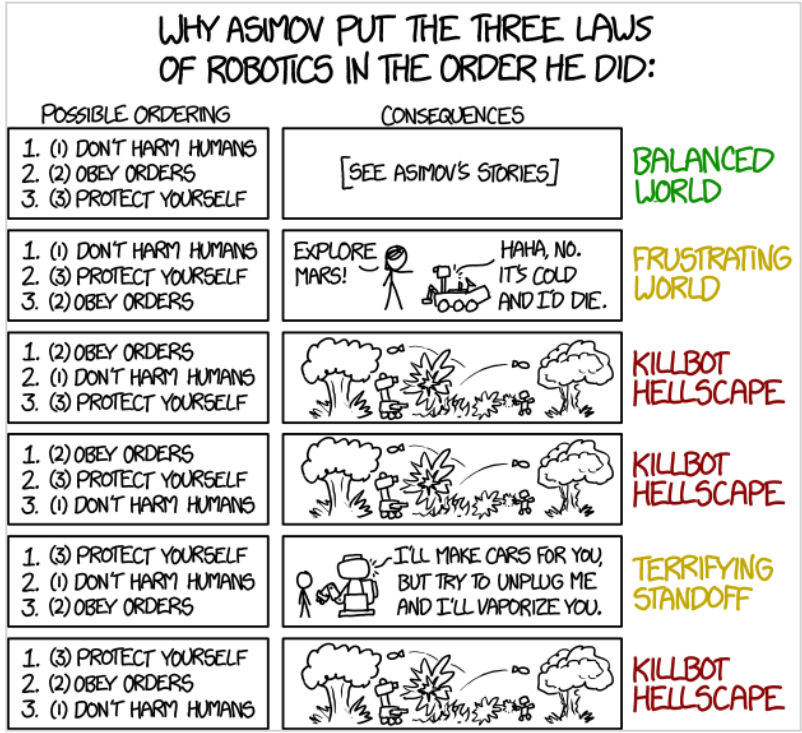
Article Talk

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