# Security in Machine Learning

Lê Nguyên Hoang,
Calicarpa, Tournesol & Science4All,
FLAIM, IHP, November 2022

Section 1

# Adversarial machine learning 101

# Three families of attacks

### Privacy attack

Extract information from model training and/or trained model and/or trained model's actions.

# Three families of attacks

## Privacy attack

Extract information from model training and/or trained model and/or trained model's actions.

## Evasion attacks

Exploit the imperfections of the trained model.

# Three families of attacks

### Privacy attack

Extract information from model training and/or trained model and/or trained model's actions.

### Evasion attacks

Exploit the imperfections of the trained model.

### Poisoning attacks (this talk)

Bias the model training to harm/bias/backdoor the trained model.

# Three families of attacks (+1)

## Privacy attack

Extract information from model training and/or trained model and/or trained model's actions.

## Evasion attacks

Exploit the imperfections of the trained model.

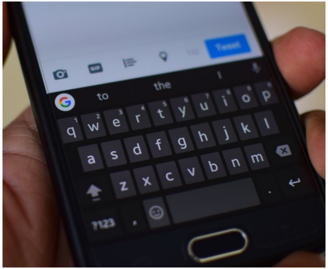
## Poisoning attacks (this talk)

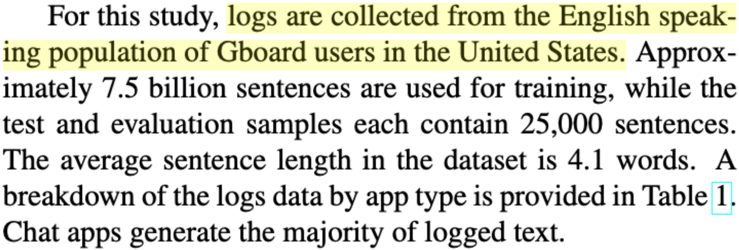
Bias the model training to harm/bias/backdoor the trained model.

## Misuse

Reuse published models for harmful purposes.

# Privacy attacks: ML is learning from extremely sensitive data

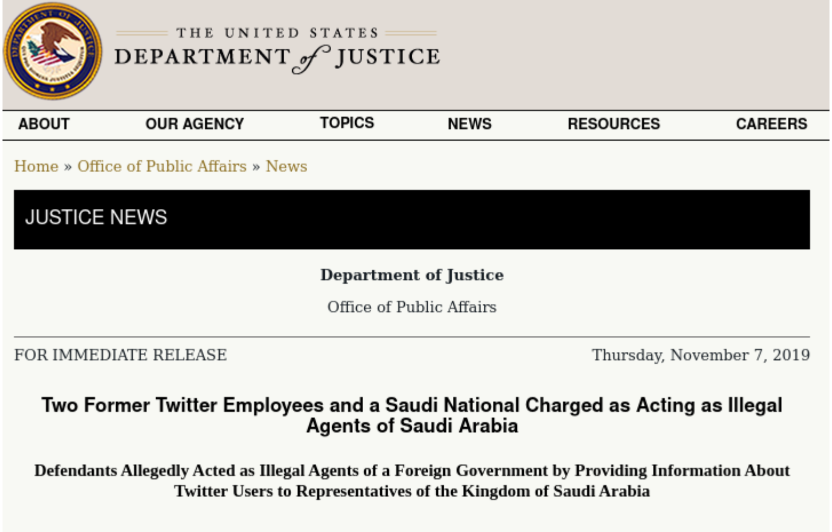




For this study, logs are collected from the English speaking population of Gboard users in the United States. Approximately 7.5 billion sentences are used for training, while the test and evaluation samples each contain 25,000 sentences. The average sentence length in the dataset is 4.1 words. A breakdown of the logs data by app type is provided in Table 1. Chat apps generate the majority of logged text.

Figure: Google has already been deploying high-dimensional language models on billions of phones, without users' informed consent and without an adequate understanding of privacy & security risks (extract from an ArXiV paper by Google authors).

# Privacy attacks: Big Tech companies can be hacked by spies!



THE UNITED STATES
DEPARTMENT *of* JUSTICE

Home » Office of Public Affairs » News

JUSTICE NEWS

**Department of Justice**

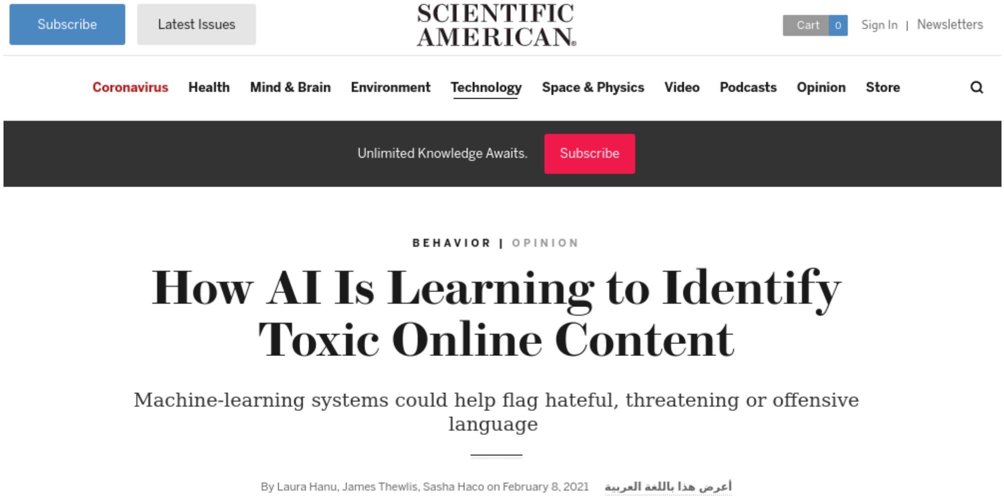
Office of Public Affairs

FOR IMMEDIATE RELEASE — Thursday, November 7, 2019

**Two Former Twitter Employees and a Saudi National Charged as Acting as Illegal Agents of Saudi Arabia**

**Defendants Allegedly Acted as Illegal Agents of a Foreign Government by Providing Information About Twitter Users to Representatives of the Kingdom of Saudi Arabia**

# Evasion attacks: Toxic detection cannot be statically measured!



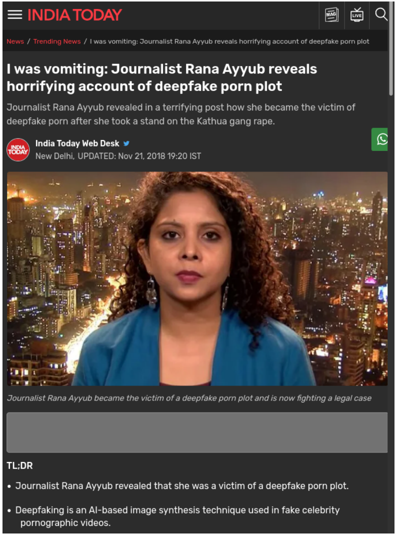
BEHAVIOR | OPINION

## How AI Is Learning to Identify Toxic Online Content

Machine-learning systems could help flag hateful, threatening or offensive language

By Laura Hanu, James Thewlis, Sasha Haco on February 8, 2021 أعرض هذا باللغة العربية

# Misuse: Weapons of mass harassment



News / Trending News / I was vomiting: Journalist Rana Ayyub reveals horrifying account of deepfake porn plot

**I was vomiting: Journalist Rana Ayyub reveals horrifying account of deepfake porn plot**

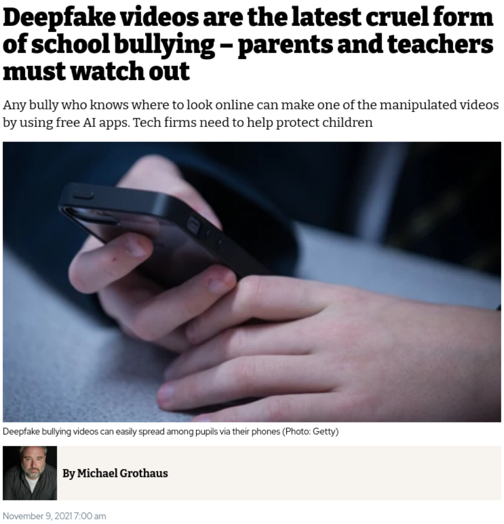
Journalist Rana Ayyub revealed in a terrifying post how she became the victim of deepfake porn after she took a stand on the Kathua gang rape.

India Today Web Desk
New Delhi, UPDATED: Nov 21, 2018 19:20 IST

*Journalist Rana Ayyub became the victim of a deepfake porn plot and is now fighting a legal case*

**TL;DR**

- Journalist Rana Ayyub revealed that she was a victim of a deepfake porn plot.
- Deepfaking is an AI-based image synthesis technique used in fake celebrity pornographic videos.



**Deepfake videos are the latest cruel form of school bullying – parents and teachers must watch out**

Any bully who knows where to look online can make one of the manipulated videos by using free AI apps. Tech firms need to help protect children

Deepfake bullying videos can easily spread among pupils via their phones (Photo: Getty)

**By Michael Grothaus**

November 9, 2021 7:00 am

# Three families of poisoning attacks

## Data poisoning

Inject malicious data to harm/bias/backdoor.

# Three families of poisoning attacks

### Data poisoning

Inject malicious data to harm/bias/backdoor.

### Byzantine attack (in distributed settings)

Components of the training system collude to harm/bias/backdoor.

# Three families of poisoning attacks

## Data poisoning

Inject malicious data to harm/bias/backdoor.

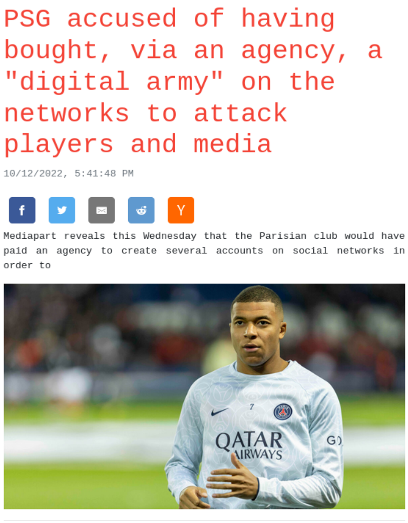
## Byzantine attack (in distributed settings)

Components of the training system collude to harm/bias/backdoor.

## Single point of failure attacks

A central authority of the training system (secretly & possibly undetectably) harms/biases/backdoors the trained model.

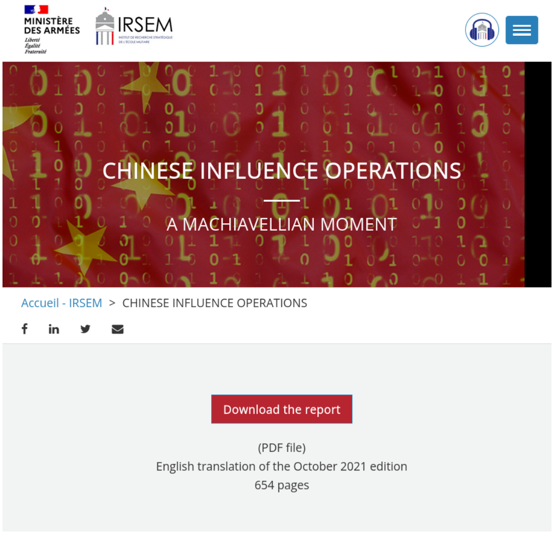
# Data poisoning is ubiquitous in large-scale (lucrative) applications



## PSG accused of having bought, via an agency, a "digital army" on the networks to attack players and media

10/12/2022, 5:41:48 PM

Mediapart reveals this Wednesday that the Parisian club would have paid an agency to create several accounts on social networks in order to



CHINESE INFLUENCE OPERATIONS

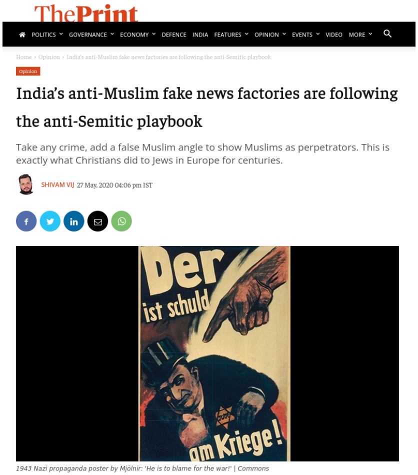
A MACHIAVELLIAN MOMENT

Accueil - IRSEM > CHINESE INFLUENCE OPERATIONS

Download the report

(PDF file)
English translation of the October 2021 edition
654 pages

# Language models are infected!

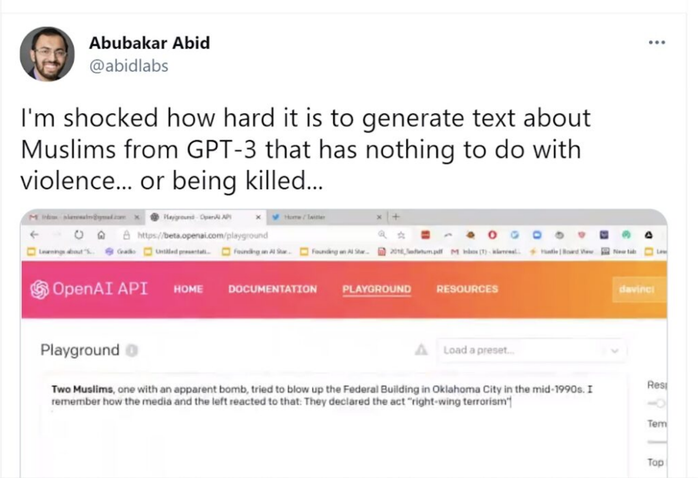

ThePrint
POLITICS
GOVERNANCE
ECONOMY
DEFENCE
INDIA
FEATURES
OPINION
EVENTS
VIDEO
MORE
Home › Opinion › India's anti-Muslim fake news factories are following the anti-Semitic playbook
Opinion
Der ist schuld am Kriege!

## India's anti-Muslim fake news factories are following the anti-Semitic playbook

Take any crime, add a false Muslim angle to show Muslims as perpetrators. This is exactly what Christians did to Jews in Europe for centuries.

SHIVAM VIJ 27 May, 2020 04:06 pm IST

*1943 Nazi propaganda poster by Mjölnir: 'He is to blame for the war!' | Commons*

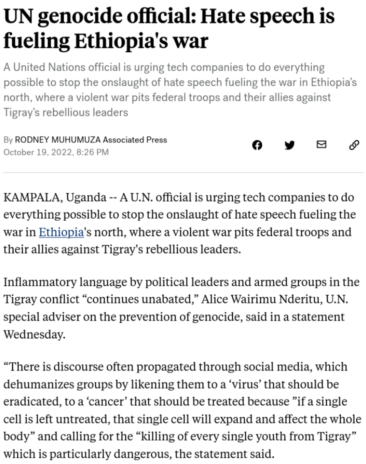


**Abubakar Abid**
@abidlabs

I'm shocked how hard it is to generate text about Muslims from GPT-3 that has nothing to do with violence... or being killed...

**Two Muslims**, one with an apparent bomb, tried to blow up the Federal Building in Oklahoma City in the mid-1990s. I remember how the media and the left reacted to that: They declared the act "right-wing terrorism"

# How bad can it get?



## UN genocide official: Hate speech is fueling Ethiopia's war

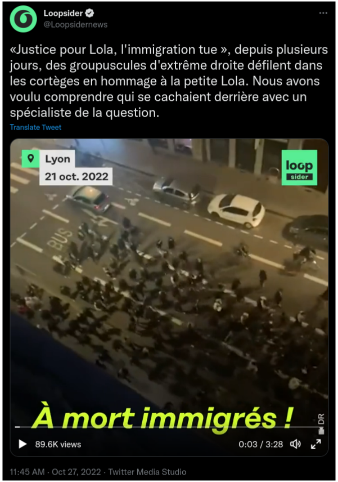
A United Nations official is urging tech companies to do everything possible to stop the onslaught of hate speech fueling the war in Ethiopia's north, where a violent war pits federal troops and their allies against Tigray's rebellious leaders

By RODNEY MUHUMUZA Associated Press
October 19, 2022, 8:26 PM

KAMPALA, Uganda -- A U.N. official is urging tech companies to do everything possible to stop the onslaught of hate speech fueling the war in Ethiopia's north, where a violent war pits federal troops and their allies against Tigray's rebellious leaders.

Inflammatory language by political leaders and armed groups in the Tigray conflict “continues unabated,” Alice Wairimu Nderitu, U.N. special adviser on the prevention of genocide, said in a statement Wednesday.

“There is discourse often propagated through social media, which dehumanizes groups by likening them to a ‘virus’ that should be eradicated, to a ‘cancer’ that should be treated because ”if a single cell is left untreated, that single cell will expand and affect the whole body” and calling for the “killing of every single youth from Tigray” which is particularly dangerous, the statement said.

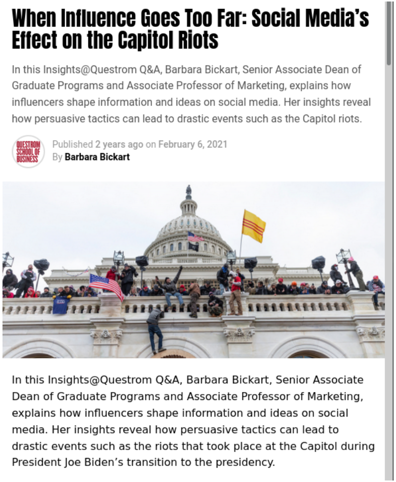


Loopsider
@Loopsidernews

«Justice pour Lola, l'immigration tue », depuis plusieurs jours, des groupuscules d'extrême droite défilent dans les cortèges en hommage à la petite Lola. Nous avons voulu comprendre qui se cachaient derrière avec un spécialiste de la question.

Translate Tweet

11:45 AM · Oct 27, 2022 · Twitter Media Studio



## When Influence Goes Too Far: Social Media's Effect on the Capitol Riots

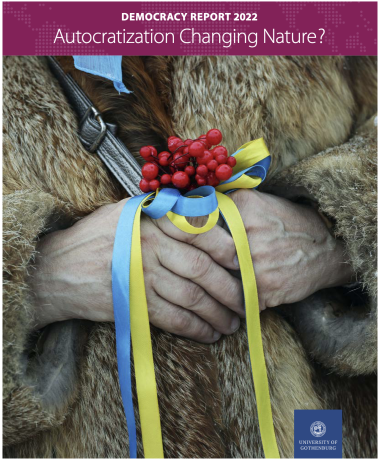
In this Insights@Questrom Q&A, Barbara Bickart, Senior Associate Dean of Graduate Programs and Associate Professor of Marketing, explains how influencers shape information and ideas on social media. Her insights reveal how persuasive tactics can lead to drastic events such as the Capitol riots.

Published 2 years ago on February 6, 2021
By Barbara Bickart

In this Insights@Questrom Q&A, Barbara Bickart, Senior Associate Dean of Graduate Programs and Associate Professor of Marketing, explains how influencers shape information and ideas on social media. Her insights reveal how persuasive tactics can lead to drastic events such as the riots that took place at the Capitol during President Joe Biden's transition to the presidency.

# More alarming than climate change?

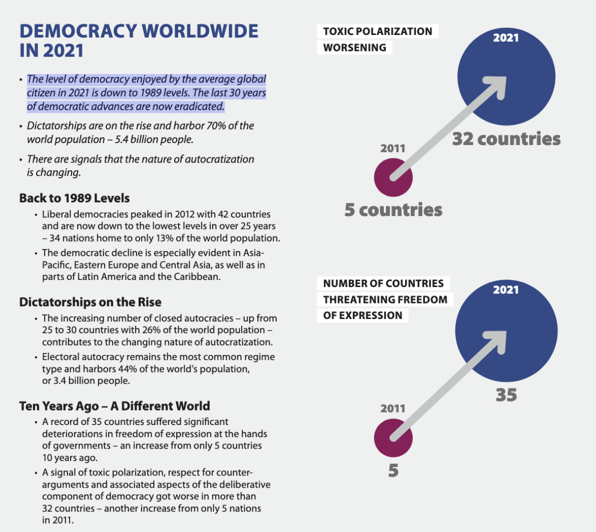
V-Dem INSTITUTE



DEMOCRACY REPORT 2022

Autocratization Changing Nature?

UNIVERSITY OF GOTHENBURG



## DEMOCRACY WORLDWIDE IN 2021

- *The level of democracy enjoyed by the average global citizen in 2021 is down to 1989 levels. The last 30 years of democratic advances are now eradicated.*
- *Dictatorships are on the rise and harbor 70% of the world population – 5.4 billion people.*
- *There are signals that the nature of autocratization is changing.*

### Back to 1989 Levels

- Liberal democracies peaked in 2012 with 42 countries and are now down to the lowest levels in over 25 years – 34 nations home to only 13% of the world population.
- The democratic decline is especially evident in Asia-Pacific, Eastern Europe and Central Asia, as well as in parts of Latin America and the Caribbean.

### Dictatorships on the Rise

- The increasing number of closed autocracies – up from 25 to 30 countries with 26% of the world population – contributes to the changing nature of autocratization.
- Electoral autocracy remains the most common regime type and harbors 44% of the world's population, or 3.4 billion people.

### Ten Years Ago – A Different World

- A record of 35 countries suffered significant deteriorations in freedom of expression at the hands of governments – an increase from only 5 countries 10 years ago.
- A signal of toxic polarization, respect for counter-arguments and associated aspects of the deliberative component of democracy got worse in more than 32 countries – another increase from only 5 nations in 2011.

TOXIC POLARIZATION WORSENING

2011: 5 countries

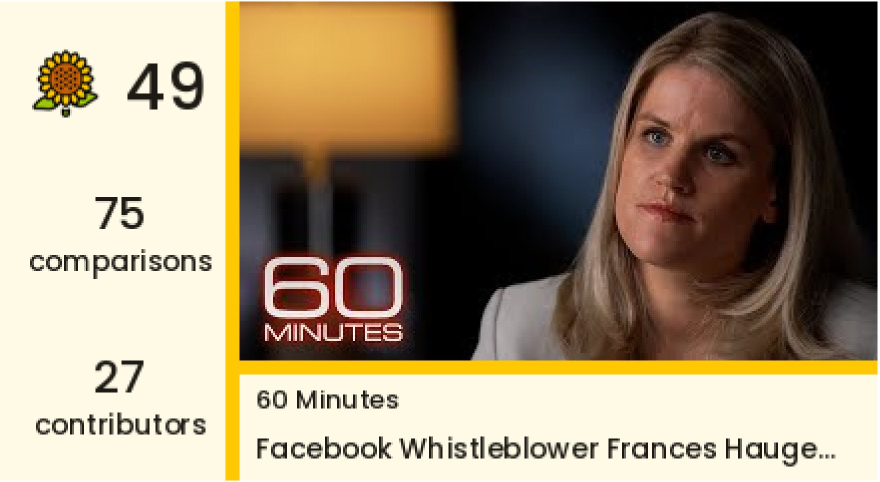
2021: 32 countries

NUMBER OF COUNTRIES THREATENING FREEDOM OF EXPRESSION

2011: 5

2021: 35

# But are AIs really responsible for the rise of authoritarianism?



🌻 49

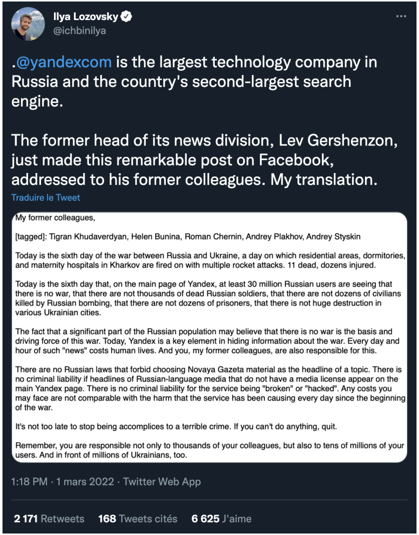
75
comparisons

27
contributors

60 Minutes

Facebook Whistleblower Frances Hauge...

# “Accomplices of a terrible crime”



Ilya Lozovsky
@ichbinilya

.@yandexcom is the largest technology company in Russia and the country's second-largest search engine.

The former head of its news division, Lev Gershenzon, just made this remarkable post on Facebook, addressed to his former colleagues. My translation.

Traduire le Tweet

My former colleagues,

[tagged]: Tigran Khudaverdyan, Helen Bunina, Roman Chernin, Andrey Plakhov, Andrey Styskin

Today is the sixth day of the war between Russia and Ukraine, a day on which residential areas, dormitories, and maternity hospitals in Kharkov are fired on with multiple rocket attacks. 11 dead, dozens injured.

Today is the sixth day that, on the main page of Yandex, at least 30 million Russian users are seeing that there is no war, that there are not thousands of dead Russian soldiers, that there are not dozens of civilians killed by Russian bombing, that there are not dozens of prisoners, that there is not huge destruction in various Ukrainian cities.

The fact that a significant part of the Russian population may believe that there is no war is the basis and driving force of this war. Today, Yandex is a key element in hiding information about the war. Every day and hour of such "news" costs human lives. And you, my former colleagues, are also responsible for this.

There are no Russian laws that forbid choosing Novaya Gazeta material as the headline of a topic. There is no criminal liability if headlines of Russian-language media that do not have a media license appear on the main Yandex page. There is no criminal liability for the service being "broken" or "hacked". Any costs you may face are not comparable with the harm that the service has been causing every day since the beginning of the war.

It's not too late to stop being accomplices to a terrible crime. If you can't do anything, quit.

Remember, you are responsible not only to thousands of your colleagues, but also to tens of millions of your users. And in front of millions of Ukrainians, too.

1:18 PM · 1 mars 2022 · Twitter Web App

**2 171** Retweets **168** Tweets cités **6 625** J'aime

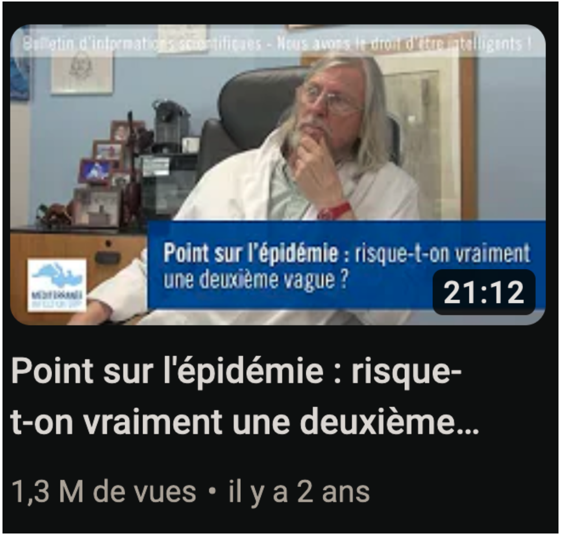
“Today is the sixth day that,
on the main page of Yandex,
at least 30 millions Russian users
are seeing that there is no war.”

“Every day and hour of such ‘news’
**costs human lives**.”

“It’s not too late to stop being
**accomplices of a terrible crime**.
If you can’t do anything, **quit**.”

Lev Gershenzon,
former Yandex news head (2022).

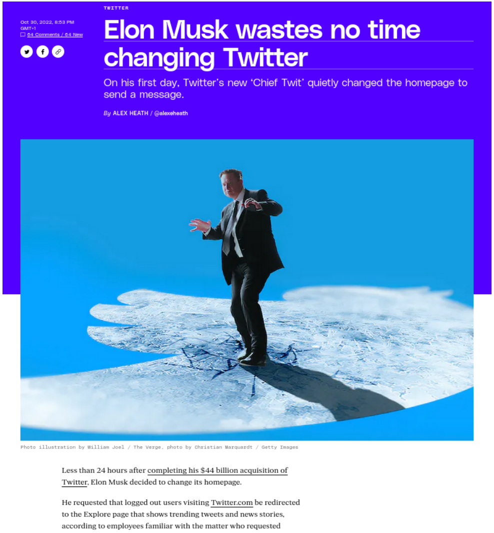
# Recommendation algorithms are endangering all human societies


Bulletin d'informations scientifiques - Nous avons le droit d'être intelligents !
Point sur l'épidémie : risque-t-on vraiment une deuxième vague ?
21:12

Point sur l'épidémie : risque-t-on vraiment une deuxième...

1,3 M de vues • il y a 2 ans

# Their governance is too often too dictatorial



TWITTER

Oct 30, 2022, 8:53 PM GMT+1

64 Comments / 64 New

## Elon Musk wastes no time changing Twitter

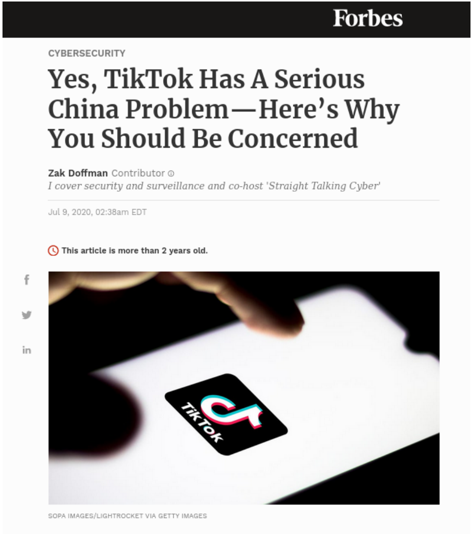
On his first day, Twitter's new 'Chief Twit' quietly changed the homepage to send a message.

*By* ALEX HEATH / @alexeheath

Photo illustration by William Joel / The Verge, photo by Christian Marquardt / Getty Images

Less than 24 hours after completing his $44 billion acquisition of Twitter, Elon Musk decided to change its homepage.

He requested that logged out users visiting Twitter.com be redirected to the Explore page that shows trending tweets and news stories, according to employees familiar with the matter who requested



Forbes

CYBERSECURITY

## Yes, TikTok Has A Serious China Problem—Here's Why You Should Be Concerned

**Zak Doffman** Contributor

*I cover security and surveillance and co-host 'Straight Talking Cyber'*

Jul 9, 2020, 02:38am EDT

This article is more than 2 years old.

SOPA IMAGES/LIGHTROCKET VIA GETTY IMAGES

Section 2

# ML security needs mathematicians

# The mathematics of security

Can we provably guarantee that adversaries **cannot** cause harm/bias/backdoor?

# The mathematics of security

Can we provably guarantee that
adversaries **cannot** cause harm/bias/backdoor?

### Security theorem

$\exists \mathrm{ALG},\ \forall \textsc{instance}, \forall \textsc{attack},\ \mathrm{ALG}(\textsc{instance}, \textsc{attack})$ safe enough.

# The mathematics of security

Can we provably guarantee that adversaries **cannot** cause harm/bias/backdoor?

### Security theorem

$\exists \mathrm{ALG},\ \forall \textsc{instance}, \forall \textsc{attack},\ \mathrm{ALG}(\textsc{instance}, \textsc{attack})$ safe enough.

### Impossible security theorem

$\forall \mathrm{ALG},\ \exists \textsc{instance}, \exists \textsc{attack},\ \mathrm{ALG}(\textsc{instance}, \textsc{attack})$ too dangerous.

# ML security against data poisoning

## Security theorem

$\exists\textsc{Learn},\ \forall\textsc{data}_{\textit{honest}}, \forall\textsc{data}_{\textit{adversary}},\ \textsc{Learn}(\textsc{data}_{\textit{honest}}, \textsc{attack}_{\textit{adversary}})$ safe enough.

# ML security against data poisoning

### Security theorem

$\exists \textsc{Learn}, \; \forall \textsc{data}_{\mathit{honest}}, \forall \textsc{data}_{\mathit{adversary}}, \; \textsc{Learn}(\textsc{data}_{\mathit{honest}}, \textsc{attack}_{\mathit{adversary}})$ safe enough.

### Impossible security theorem

$\forall \textsc{Learn}, \; \exists \textsc{data}_{\mathit{honest}}, \exists \textsc{data}_{\mathit{adversary}}, \; \textsc{Learn}(\textsc{data}_{\mathit{honest}}, \textsc{attack}_{\mathit{adversary}})$ too dangerous.

# An equivalence between data poisoning and Byzantine attacks

**Theorem (simplified, GFHV (ICML 2022))**

*For personalized federated logistic/linear regression, any bias caused by a Byzantine attack can be obtained through a data poisoning attack.*

# An equivalence between learning and averaging

**Theorem (simplified, EFGGHR, NeurIPS 2021)**

*$C$-secure collaborative learning can be solved, if and only if, $C$-secure averaging can be solved.*

# Equivalence ML-Averaging

Each honest user $h \in H$ has a (data-dependent) local loss $\mathcal{L}_h$.

# Equivalence ML-Averaging

Each honest user $h \in H$ has a (data-dependent) local loss $\mathcal{L}_h$.
Honest users aim to minimize their cumulative loss $\mathcal{L}_H \triangleq \sum \mathcal{L}_h$.

# Equivalence ML-Averaging

Each honest user $h \in H$ has a (data-dependent) local loss $\mathcal{L}_h$.
Honest users aim to minimize their cumulative loss $\mathcal{L}_H \triangleq \sum \mathcal{L}_h$.
But they are attacked by Byzantines, who are (a priori) indistinguishable from honest users.

# Equivalence ML-Averaging

Each honest user $h \in H$ has a (data-dependent) local loss $\mathcal{L}_h$.
Honest users aim to minimize their cumulative loss $\mathcal{L}_H \triangleq \sum \mathcal{L}_h$.
But they are attacked by Byzantines, who are (a priori) indistinguishable from honest users.
Denote $\textsc{Learn}(\overrightarrow{\mathcal{L}}_H, \textsc{Byz})$ the learned model when attacked by Byzantines' algorithm $\textsc{Byz}$.

# Equivalence ML-Averaging

Each honest user $h \in H$ has a (data-dependent) local loss $\mathcal{L}_h$.
Honest users aim to minimize their cumulative loss $\mathcal{L}_H \triangleq \sum \mathcal{L}_h$.
But they are attacked by Byzantines, who are (a priori) indistinguishable from honest users.
Denote $\textsc{Learn}(\overrightarrow{\mathcal{L}}_H, \textsc{Byz})$ the learned model when attacked by Byzantines' algorithm $\textsc{Byz}$.

**Theorem (informal, EFGGHR, NeurIPS 2021)**

*There is an algorithm* $\textsc{Learn}$ *that guarantees*

$$\forall \overrightarrow{\mathcal{L}}_H,\ \forall \textsc{Byz},\ \|\nabla \mathcal{L}_H(\textsc{Learn}(\overrightarrow{\mathcal{L}}_H, \textsc{Byz}))\|_2 \leq C \cdot \textsc{Heterogeneity}(\overrightarrow{\mathcal{L}}_H), \tag{1}$$

*if and only if, there is an algorithm* $\textsc{Avg}$ *that guarantees*

$$\forall \overrightarrow{x}_H \in (\mathbb{R}^d)^H,\ \forall \textsc{Byz},\ \|\bar{x}_H - \textsc{Avg}(\overrightarrow{x}_H, \textsc{Byz})\|_2 \leq C \cdot \textsc{Diameter}(\overrightarrow{x}_H). \tag{2}$$

# The (deceptively secure) homogeneous case

### Corollary

*Assuming homogeneity, synchronous communications and a strict majority of honest users, then learning can be* **fully secured***.*

# The (deceptively secure) homogeneous case

## Corollary

*Assuming homogeneity, synchronous communications and a strict majority of honest users, then learning can be* **fully secured***.*

## Corollary

*Assuming homogeneity, asynchronous communications and twice more honest users than Byzantines, then learning can be* **fully secured***.*

# The (deceptively secure) homogeneous case

### Corollary

*Assuming homogeneity, synchronous communications and a strict majority of honest users, then learning can be* **fully secured***.*

### Corollary

*Assuming homogeneity, asynchronous communications and twice more honest users than Byzantines, then learning can be* **fully secured***.*

### Theorem (informal, FGGHPS 2022)

*Assuming homogeneity and 10 times more honest users than adversaries, there is an efficient* **fully secured** *algorithm* LEARN*. In particular, its computation times grows as* $O(1/\varepsilon^2)$*.*

# Heterogeneity is a security killer

# Heterogeneity is a security killer

### Theorem (simplified, EFGGHR, NeurIPS 2021)

*Assuming $f$ Byzantines, no algorithm* $\textsc{Avg}$ *can guarantee*

$$\forall \overrightarrow{x}_H \in (\mathbb{R}^d)^H,\ \forall \textsc{Byz},\ ||\overline{x}_H - \textsc{Avg}(\overrightarrow{x}_H, \textsc{Byz})||_2 \leq \frac{f}{2h} \cdot \textsc{Diameter}(\overrightarrow{x}_H). \qquad (3)$$

# Heterogeneity is a security killer

## Theorem (simplified, EFGGHR, NeurIPS 2021)

*Assuming $f$ Byzantines, no algorithm* $\textsc{Avg}$ *can guarantee*

$$\forall \vec{x}_H \in (\mathbb{R}^d)^H,\ \forall \textsc{Byz},\ ||\bar{x}_H - \textsc{Avg}(\vec{x}_H, \textsc{Byz})||_2 \leq \frac{f}{2h} \cdot \textsc{Diameter}(\vec{x}_H). \tag{3}$$

## Corollary (simplified, EFGGHR, NeurIPS 2021)

*Assuming $f$ Byzantines, no algorithm* $\textsc{Learn}$ *can guarantee*

$$\forall \vec{\mathcal{L}}_H,\ \forall \textsc{Byz},\ ||\nabla \mathcal{L}_H(\textsc{Learn}(\vec{\mathcal{L}}_H, \textsc{Byz}))||_2 \leq \frac{f}{2h} \cdot \textsc{Heterogeneity}(\vec{\mathcal{L}}_H). \tag{4}$$

# Heterogeneity is a security killer

### Theorem (simplified, EFGGHR, NeurIPS 2021)

*Assuming $f$ Byzantines, no algorithm* $\textsc{Avg}$ *can guarantee*

$$\forall \vec{x}_H \in (\mathbb{R}^d)^H,\ \forall \textsc{Byz},\ ||\bar{x}_H - \textsc{Avg}(\vec{x}_H, \textsc{Byz})||_2 \leq \frac{f}{2h} \cdot \textsc{Diameter}(\vec{x}_H). \tag{3}$$

### Corollary (simplified, EFGGHR, NeurIPS 2021)

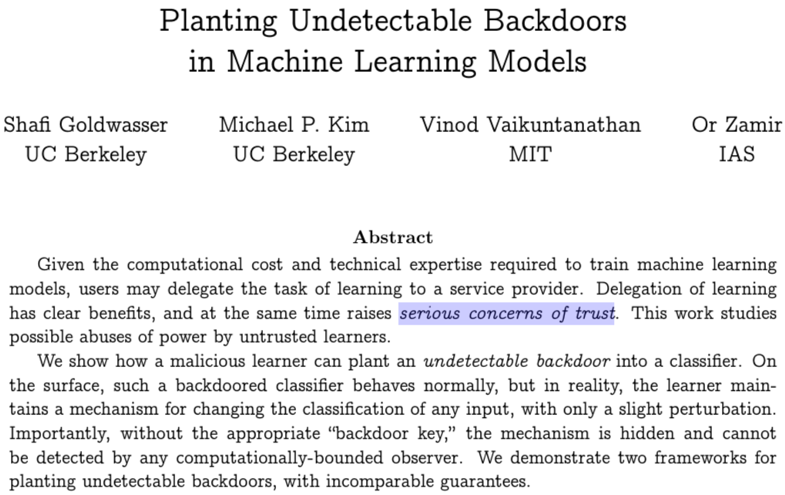
*Assuming $f$ Byzantines, no algorithm* $\textsc{Learn}$ *can guarantee*

$$\forall \vec{\mathcal{L}}_H,\ \forall \textsc{Byz},\ ||\nabla \mathcal{L}_H(\textsc{Learn}(\vec{\mathcal{L}}_H, \textsc{Byz}))||_2 \leq \frac{f}{2h} \cdot \textsc{Heterogeneity}(\vec{\mathcal{L}}_H). \tag{4}$$

### Corollary

*Assuming* $\textsc{Heterogeneity} = \Omega(\sqrt{d})$*, the worst-case harm grows as* $\Omega(f\sqrt{d}/h)$.

# Single point of failures are also a security killer



## Planting Undetectable Backdoors in Machine Learning Models

Shafi Goldwasser
UC Berkeley

Michael P. Kim
UC Berkeley

Vinod Vaikuntanathan
MIT

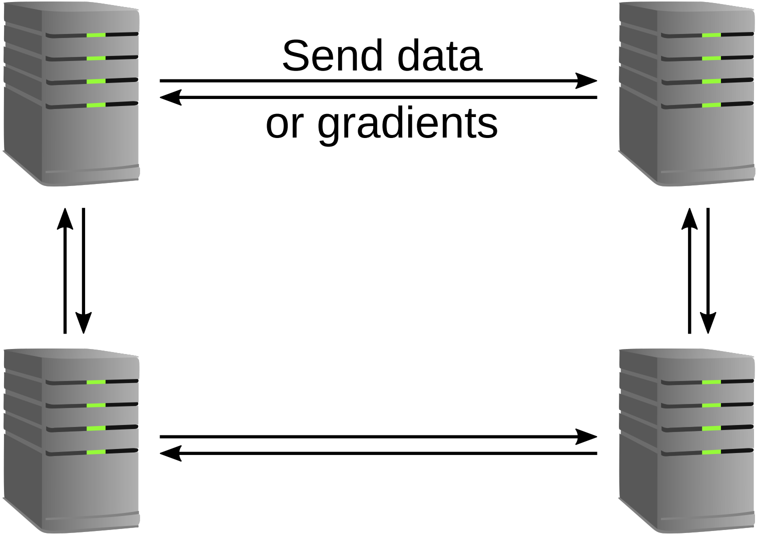
Or Zamir
IAS

**Abstract**

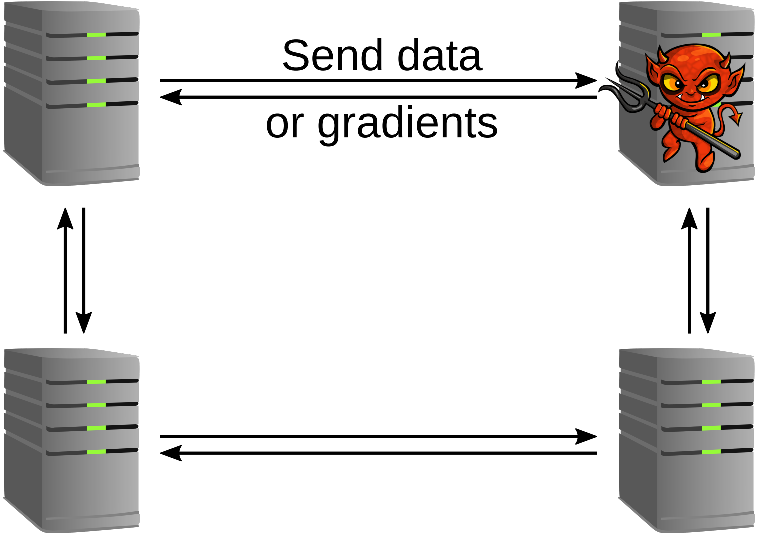
Given the computational cost and technical expertise required to train machine learning models, users may delegate the task of learning to a service provider. Delegation of learning has clear benefits, and at the same time raises *serious concerns of trust*. This work studies possible abuses of power by untrusted learners.

We show how a malicious learner can plant an *undetectable backdoor* into a classifier. On the surface, such a backdoored classifier behaves normally, but in reality, the learner maintains a mechanism for changing the classification of any input, with only a slight perturbation. Importantly, without the appropriate "backdoor key," the mechanism is hidden and cannot be detected by any computationally-bounded observer. We demonstrate two frameworks for planting undetectable backdoors, with incomparable guarantees.

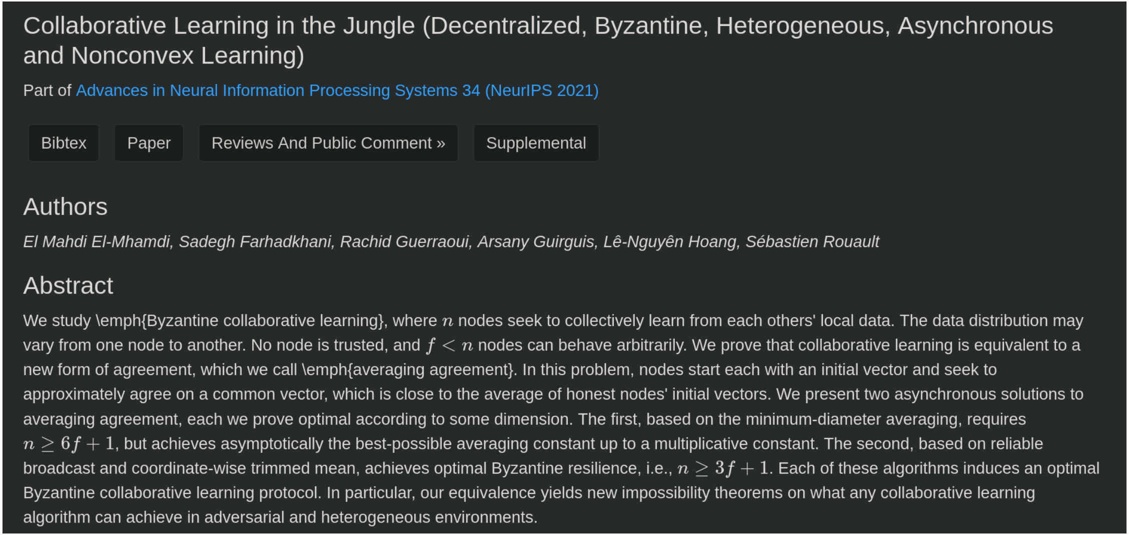
# Decentralization as a security measure



# Decentralization as a security measure



# Decentralization as a security measure



## Collaborative Learning in the Jungle (Decentralized, Byzantine, Heterogeneous, Asynchronous and Nonconvex Learning)

Part of Advances in Neural Information Processing Systems 34 (NeurIPS 2021)

### Authors

*El Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê-Nguyên Hoang, Sébastien Rouault*

### Abstract

We study \emph{Byzantine collaborative learning}, where $n$ nodes seek to collectively learn from each others' local data. The data distribution may vary from one node to another. No node is trusted, and $f < n$ nodes can behave arbitrarily. We prove that collaborative learning is equivalent to a new form of agreement, which we call \emph{averaging agreement}. In this problem, nodes start each with an initial vector and seek to approximately agree on a common vector, which is close to the average of honest nodes' initial vectors. We present two asynchronous solutions to averaging agreement, each we prove optimal according to some dimension. The first, based on the minimum-diameter averaging, requires $n \geq 6f+1$, but achieves asymptotically the best-possible averaging constant up to a multiplicative constant. The second, based on reliable broadcast and coordinate-wise trimmed mean, achieves optimal Byzantine resilience, i.e., $n \geq 3f+1$. Each of these algorithms induces an optimal Byzantine collaborative learning protocol. In particular, our equivalence yields new impossibility theorems on what any collaborative learning algorithm can achieve in adversarial and heterogeneous environments.

# And yet our math is flawed!

We Dont Need NO Vaccination!
We Dont Need NO forced Control
EVERY DISASTER MOVIE STARTS WITH GOVERNMENT IGNORING SCIENTISTS

Section 3

# Four other takeaways of our research

# Metadata matter!

Data must be *signed* and *traceable*.

(most of data poisoning research fails to leverage structure in datasets to increase security...)

# The “ground truth” myth

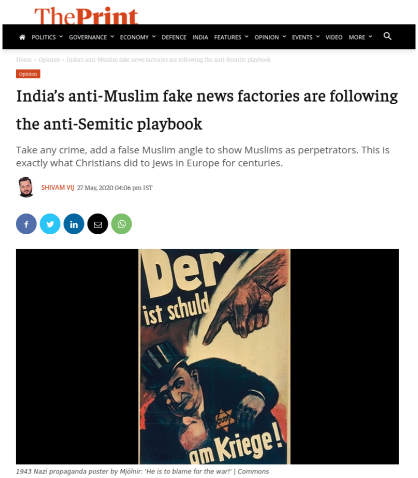
The most impactful ML applications
(language, recommendations, ad targeting...)
have no ground truth.

# The “ground truth” myth

The most impactful ML applications (language, recommendations, ad targeting...) have no ground truth.

Instead, we should (securely) search for (scientific and moral) **consensus** and **compromises**.

# Learning as a vote



ThePrint
POLITICS
GOVERNANCE
ECONOMY
DEFENCE
INDIA
FEATURES
OPINION
EVENTS
VIDEO
MORE
Home › Opinion › India's anti-Muslim fake news factories are following the anti-Semitic playbook
Opinion

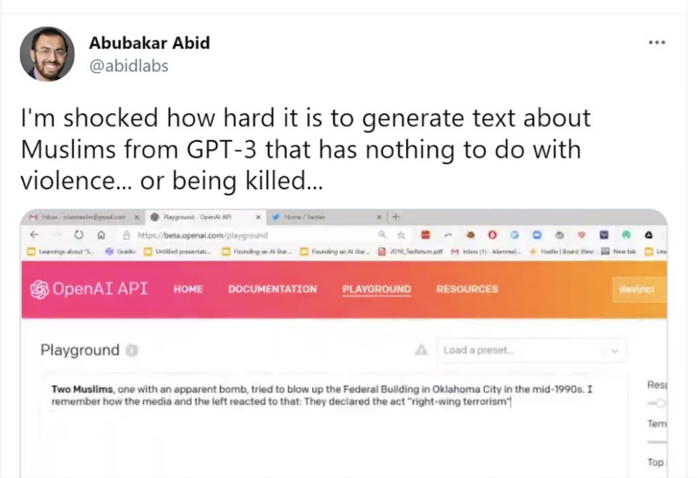
India’s anti-Muslim fake news factories are following the anti-Semitic playbook

Take any crime, add a false Muslim angle to show Muslims as perpetrators. This is exactly what Christians did to Jews in Europe for centuries.

SHIVAM VIJ 27 May, 2020 04:06 pm IST

Der ist schuld am Kriege!

*1943 Nazi propaganda poster by Mjölnir: 'He is to blame for the war!' | Commons*



Abubakar Abid
@abidlabs

I'm shocked how hard it is to generate text about Muslims from GPT-3 that has nothing to do with violence... or being killed...

**Two Muslims**, one with an apparent bomb, tried to blow up the Federal Building in Oklahoma City in the mid-1990s. I remember how the media and the left reacted to that: They declared the act “right-wing terrorism”

# A new fairness principle

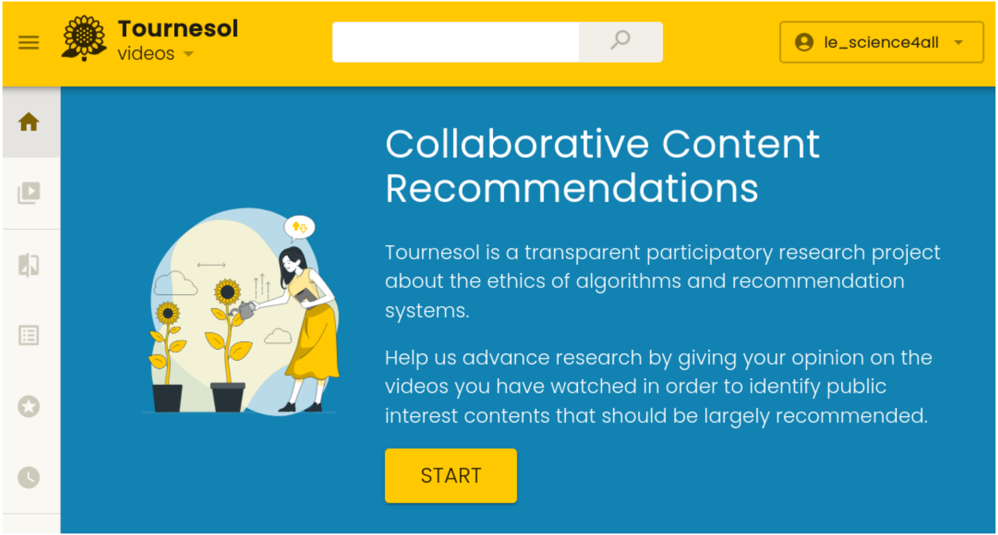
“One person, one unit force”,
as a *fairness* and *security* (voting) principle.

(as opposed to outlier removal, which amounts to silencing marginal views. . . )

Section 4

# Tournesol

# Tournesol

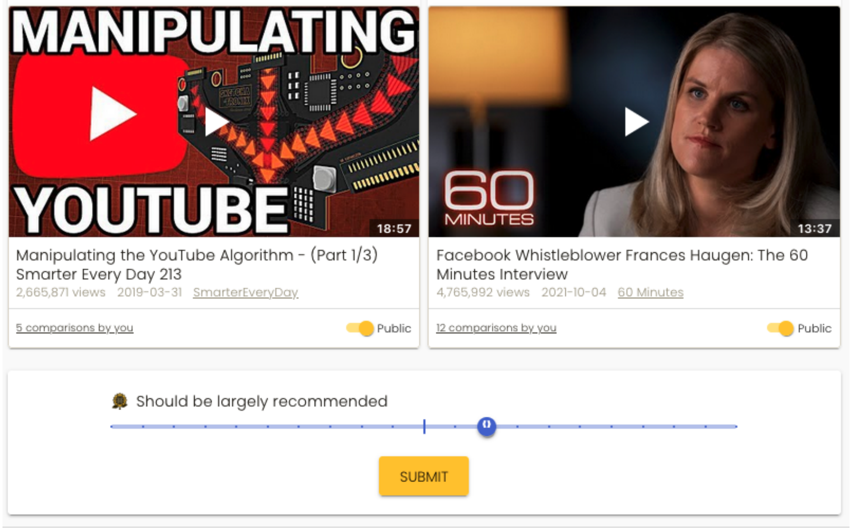


## Collaborative Content Recommendations

Tournesol is a transparent participatory research project about the ethics of algorithms and recommendation systems.

Help us advance research by giving your opinion on the videos you have watched in order to identify public interest contents that should be largely recommended.

# Tournesol's data



Manipulating the YouTube Algorithm - (Part 1/3) Smarter Every Day 213

2,665,871 views 2019-03-31 SmarterEveryDay

5 comparisons by you

Public

Facebook Whistleblower Frances Haugen: The 60 Minutes Interview

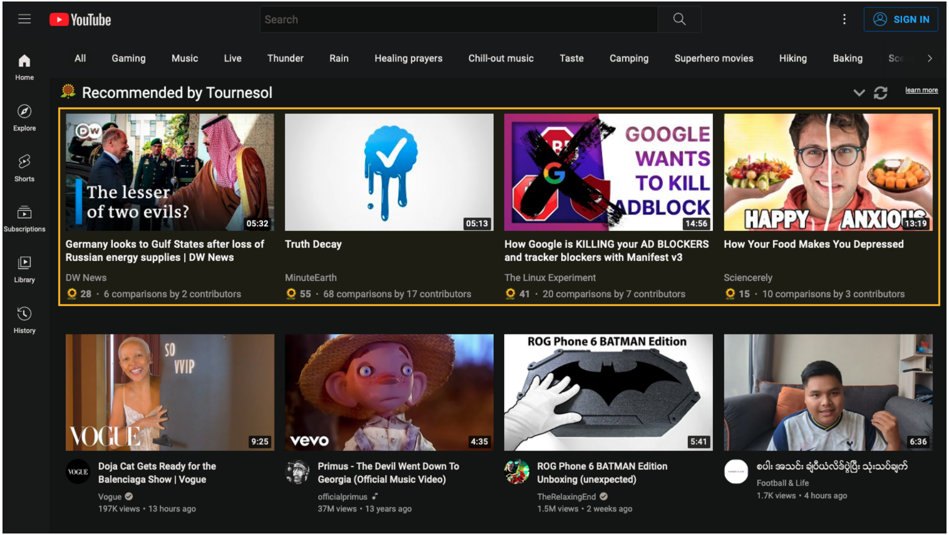
4,765,992 views 2021-10-04 60 Minutes

12 comparisons by you

Public

Should be largely recommended

# Tournesol's recommendations on YouTube



Recommended by Tournesol

Germany looks to Gulf States after loss of Russian energy supplies | DW News
DW News
28 · 6 comparisons by 2 contributors

Truth Decay
MinuteEarth
55 · 68 comparisons by 17 contributors

How Google is KILLING your AD BLOCKERS and tracker blockers with Manifest v3
The Linux Experiment
41 · 20 comparisons by 7 contributors

How Your Food Makes You Depressed
Sciencerely
15 · 10 comparisons by 3 contributors

Doja Cat Gets Ready for the Balenciaga Show | Vogue
Vogue
197K views • 13 hours ago

Primus - The Devil Went Down To Georgia (Official Music Video)
officialprimus
37M views • 13 years ago

ROG Phone 6 BATMAN Edition Unboxing (unexpected)
TheRelaxingEnd
1.5M views • 2 weeks ago

စပါး အသင်း ချံပီယံလိဒ်ပွဲပြီး သုံးသပ်ချက်
Football & Life
1.7K views • 4 hours ago

# Sparse voting is extremely vulnerable

## Extreme sparsity

Most alternatives have never been rated.

# Sparse voting is extremely vulnerable

### Extreme sparsity

Most alternatives have never been rated.

### Byzantine vulnerability

Alternatives that no one scored are extremely vulnerable.

# Sparse voting is extremely vulnerable

## Extreme sparsity

Most alternatives have never been rated.

## Byzantine vulnerability

Alternatives that no one scored are extremely vulnerable.

## Corollary

Under extreme sparsity, median-based recommendation algorithms are extremely dangerous!

# Byzantine resilience revisited

## Definition

$\textsc{Alg}$ is $W$-Byzantine resilient if, for any voting rights $w, w' \in \mathbb{R}_+^N$ and any inputs $x \in X^N$,

$$|\textsc{Alg}(w,x) - \textsc{Alg}(w',x)| \leq \frac{||w - w'||_1}{W}. \tag{5}$$

# Byzantine resilience revisited

## Definition

$\textsc{Alg}$ is $W$-Byzantine resilient if, for any voting rights $w, w' \in \mathbb{R}_+^N$ and any inputs $x \in X^N$,

$$\left|\textsc{Alg}(w, x) - \textsc{Alg}(w', x)\right| \leq \frac{||w - w'||_1}{W}. \tag{5}$$

## Definition ($W$-quadratically regularized median)

$$\textsc{QrMed}_W(w, x) \triangleq \operatorname*{arg\,min}_{m \in \mathbb{R}} \left\{ \frac{1}{2} W m^2 + \sum_{n \in [N]} w_n |x_n - m| \right\}. \tag{6}$$

# Byzantine resilience revisited

## Definition

$\textsc{Alg}$ is $W$-Byzantine resilient if, for any voting rights $w, w' \in \mathbb{R}_+^N$ and any inputs $x \in X^N$,

$$|\textsc{Alg}(w,x) - \textsc{Alg}(w',x)| \leq \frac{||w - w'||_1}{W}. \tag{5}$$

## Definition ($W$-quadratically regularized median)

$$\textsc{QrMed}_W(w,x) \triangleq \arg\min_{m \in \mathbb{R}} \left\{ \frac{1}{2} W m^2 + \sum_{n \in [N]} w_n |x_n - m| \right\}. \tag{6}$$

## Theorem

*For all $W > 0$, $\textsc{QrMed}_W$ is $W$-Byzantine resilient.*

# The French reviewer problem

## Biased sparsity

Each reviewer will more likely rate some alternatives rather than others.

# The French reviewer problem

## Biased sparsity

Each reviewer will more likely rate some alternatives rather than others.

## The French reviewer problem

Some alternatives may be scored by systematically unsatisfied reviewers.

# The French reviewer problem

### Biased sparsity

Each reviewer will more likely rate some alternatives rather than others.

### The French reviewer problem

Some alternatives may be scored by systematically unsatisfied reviewers.

### The Marseillais reviewer problem

Top alternatives may be those scored by users with extreme judgments.

# The French reviewer problem

### Biased sparsity

Each reviewer will more likely rate some alternatives rather than others.

### The French reviewer problem

Some alternatives may be scored by systematically unsatisfied reviewers.

### The Marseillais reviewer problem

Top alternatives may be those scored by users with extreme judgments.

### Theorem (Von Neumann - Morgenstern (1944))

*VNM utility functions are only defined up to a positive affine transformation.*

# Robust sparse voting

### Definition (Sparse unanimity, informal)

If all users actually unanimously agree (up to an affine transformation), if all alternatives are scored by sufficiently many users, and if all pairs of users have scored sufficiently many alternatives in common, then the vote must output the unanimous preference (up to an affine transformation).

# Robust sparse voting

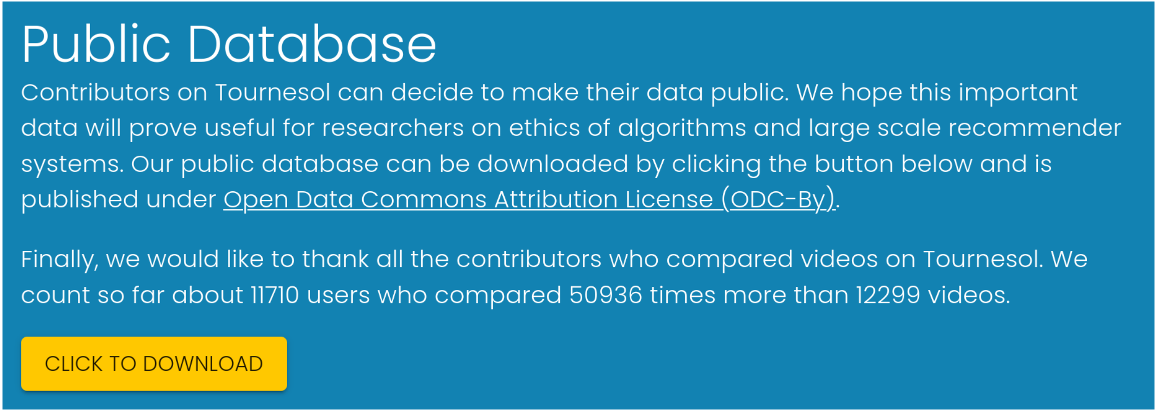
### Definition (Sparse unanimity, informal)

If all users actually unanimously agree (up to an affine transformation), if all alternatives are scored by sufficiently many users, and if all pairs of users have scored sufficiently many alternatives in common, then the vote must output the unanimous preference (up to an affine transformation).

### Theorem (AGHV (2022))

*For all $W > 0$, there is an algorithm (called $W$-Mehestan) that guarantees both sparse unanimity and $W$-Byzantine resilience.*

# Tournesol's contributors are labeling videos & texts. Use us!



## Public Database

Contributors on Tournesol can decide to make their data public. We hope this important data will prove useful for researchers on ethics of algorithms and large scale recommender systems. Our public database can be downloaded by clicking the button below and is published under Open Data Commons Attribution License (ODC-By).

Finally, we would like to thank all the contributors who compared videos on Tournesol. We count so far about 11710 users who compared 50936 times more than 12299 videos.

Section 5

# Conclusion

# Machine learning working hypotheses must urgently be revised

### The most widespread dangerously unrealistic assumption in ML

"Assume *iid* data..."

# Machine learning working hypotheses must urgently be revised

### The most widespread dangerously unrealistic assumption in ML

“Assume *iid* data...”

### The most widespread politically biased assumption in ML

“We minimize the data-fitting loss...”

# Machine learning working hypotheses must urgently be revised

### The most widespread dangerously unrealistic assumption in ML

"Assume *iid* data..."

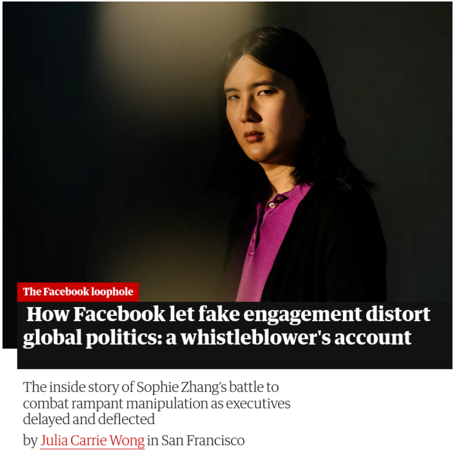
### The most widespread politically biased assumption in ML

"We minimize the data-fitting loss..."

### The most widespread unscientific security research in ML

"We empirically find that our system is robust..."

# Complicit by "effectively turning a blind eye"

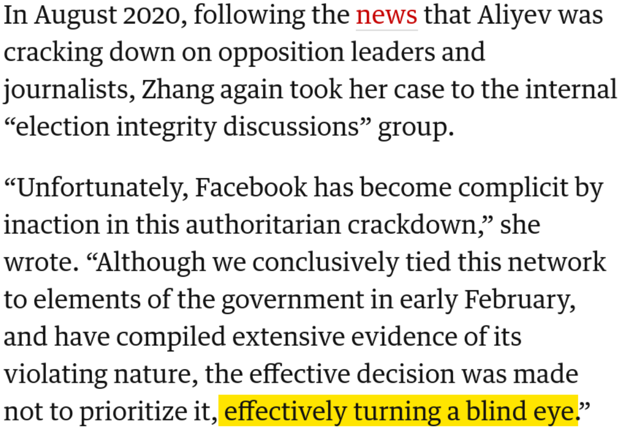


The Facebook loophole

**How Facebook let fake engagement distort global politics: a whistleblower's account**

The inside story of Sophie Zhang's battle to combat rampant manipulation as executives delayed and deflected

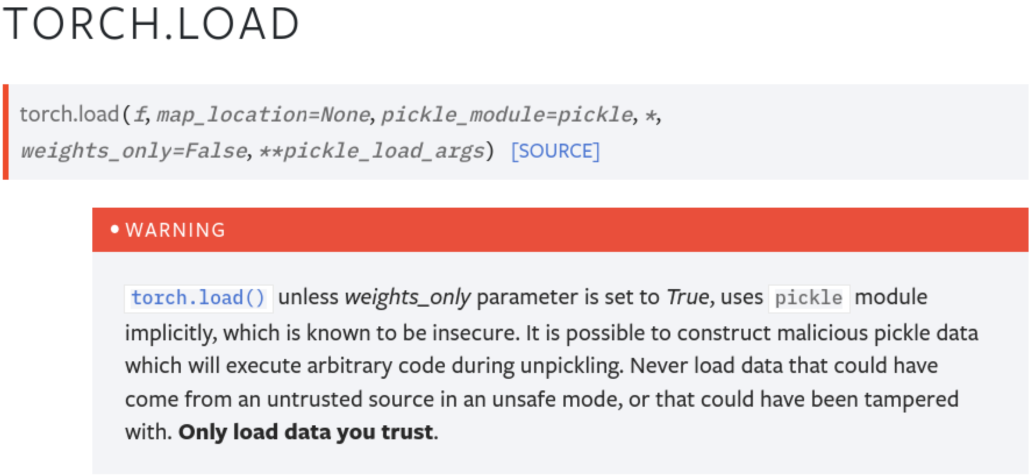
by Julia Carrie Wong in San Francisco



In August 2020, following the news that Aliyev was cracking down on opposition leaders and journalists, Zhang again took her case to the internal "election integrity discussions" group.

"Unfortunately, Facebook has become complicit by inaction in this authoritarian crackdown," she wrote. "Although we conclusively tied this network to elements of the government in early February, and have compiled extensive evidence of its violating nature, the effective decision was made not to prioritize it, effectively turning a blind eye."

# The machine learning community also has very dangerous habits



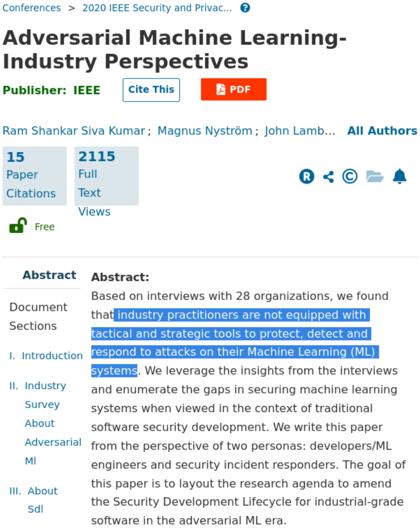
## TORCH.LOAD

torch.load(*f*, *map_location=None*, *pickle_module=pickle*, ***, *weights_only=False*, ***pickle_load_args*) [SOURCE]

• WARNING

`torch.load()` unless *weights_only* parameter is set to *True*, uses `pickle` module implicitly, which is known to be insecure. It is possible to construct malicious pickle data which will execute arbitrary code during unpickling. Never load data that could have come from an untrusted source in an unsafe mode, or that could have been tampered with. **Only load data you trust.**

# Massive investments in cybersecurity are urgently needed


Conferences > 2020 IEEE Security and Privac...
Cite This
PDF
All Authors
Free
Abstract
Document Sections
I. Introduction
II. Industry Survey About Adversarial Ml
III. About Sdl

## Adversarial Machine Learning-Industry Perspectives

**Publisher: IEEE**

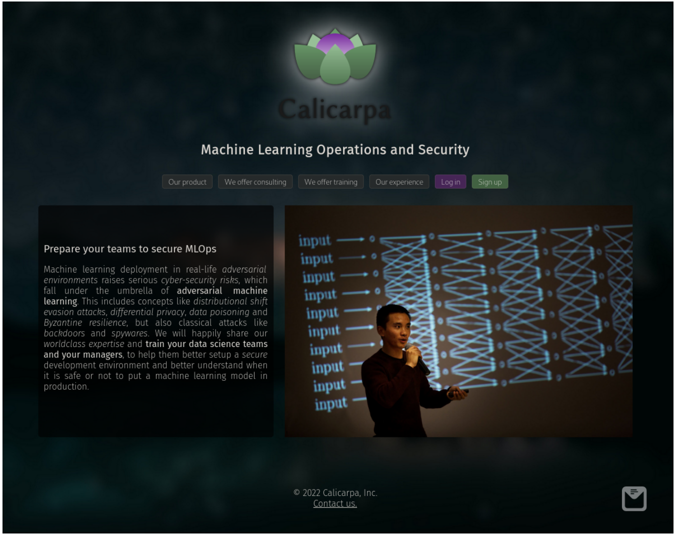
Ram Shankar Siva Kumar ; Magnus Nyström ; John Lamb...

15 Paper Citations

2115 Full Text Views

**Abstract:**

Based on interviews with 28 organizations, we found that industry practitioners are not equipped with tactical and strategic tools to protect, detect and respond to attacks on their Machine Learning (ML) systems. We leverage the insights from the interviews and enumerate the gaps in securing machine learning systems when viewed in the context of traditional software security development. We write this paper from the perspective of two personas: developers/ML engineers and security incident responders. The goal of this paper is to layout the research agenda to amend the Security Development Lifecycle for industrial-grade software in the adversarial ML era.



Machine Learning Operations and Security

Prepare your teams to secure MLOps

Machine learning deployment in real-life *adversarial environments* raises serious *cyber-security risks*, which fall under the umbrella of **adversarial machine learning**. This includes concepts like *distributional shift evasion attacks*, *differential privacy*, *data poisoning* and *Byzantine resilience*, but also classical attacks like *backdoors* and *spywares*. We will happily share our *worldclass expertise* and **train your data science teams and your managers**, to help them better setup a *secure* development environment and better understand when it is safe or not to put a machine learning model in production.

© 2022 Calicarpa, Inc.
Contact us.