Smashing the ML Stack for Fun and Lawsuits

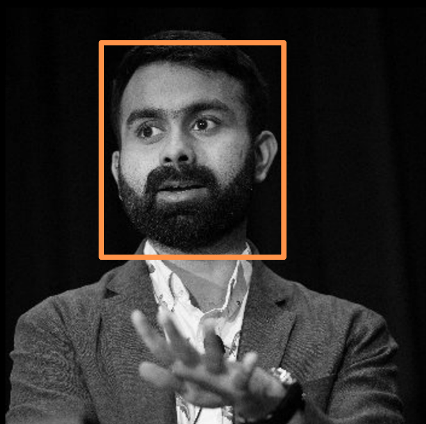
Ram Shankar Siva Kumar, Azure Trustworthy ML

Kendra Albert, Harvard Law School

@ram_ssk @kendraserra

black hat USA 2021

About us



Ram Shankar Siva Kumar
(he/him)
Microsoft

Kendra Albert
(they/them)
Harvard Law School

Jon Penney
(he/him)
York University

Bruce Schneier
(he/him)
Harvard Kennedy School

Adversarial machine learning introduces novel legal problems for security researchers…

…but what's clear is that for those who provide ML systems, technical defenses can help shore up legal claims

Agenda

01 Attacking machine learning systems

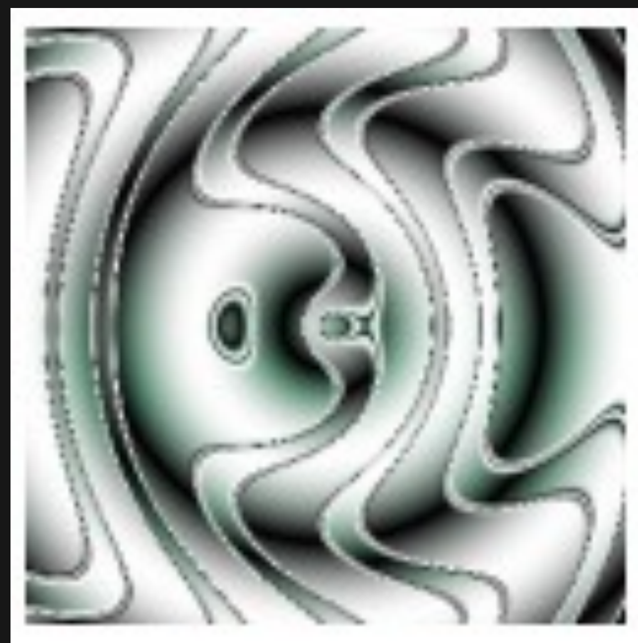
02 Legal implications for AI researchers

03 Way forward

01 Adversarial machine learning

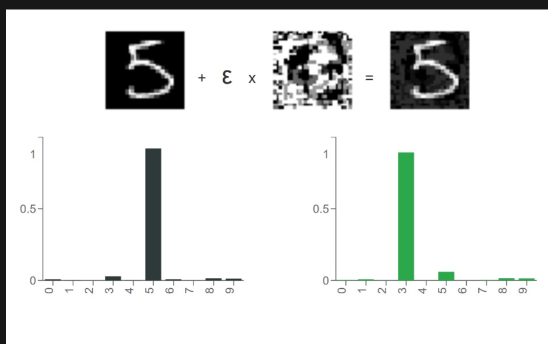
A brief quiz...

@ram_ssk @kendraserra



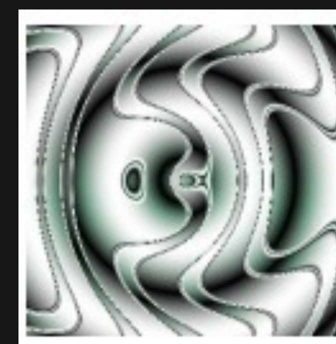
Congratulations,
you’re 100% human!

What a ML system sees…



Initial class: 5 Predicted class: 3

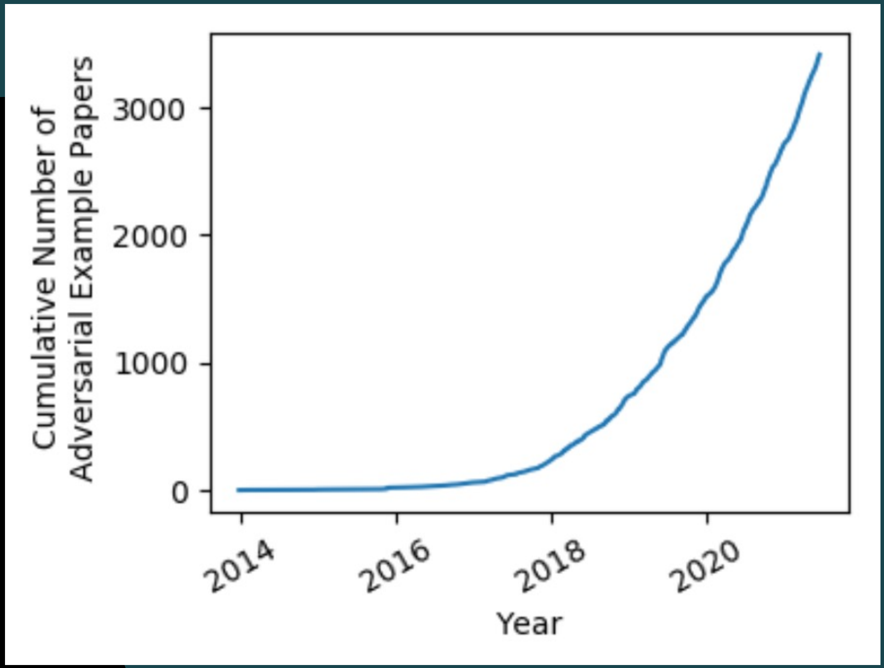
"Bird" "Orangutan"



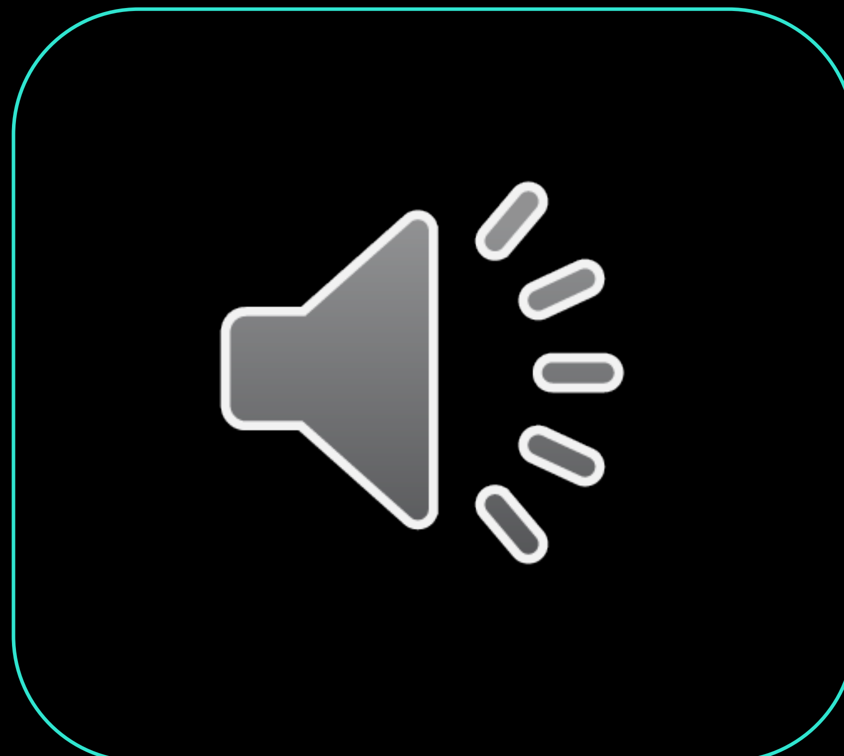
"Vacuum cleaner"

Source: https://arxiv.org/abs/1809.08352 | https://arxiv.org/pdf/1412.1897.pdf

Boom in adversarial ML research

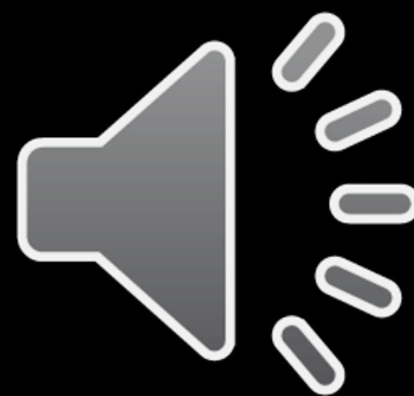


Source: Nicolas Carlini - https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html



Source: https://arxiv.org/abs/1801.01944

@ram_ssk @kendraserra



Doesn’t transcribe to anything

Source: https://arxiv.org/abs/1801.01944

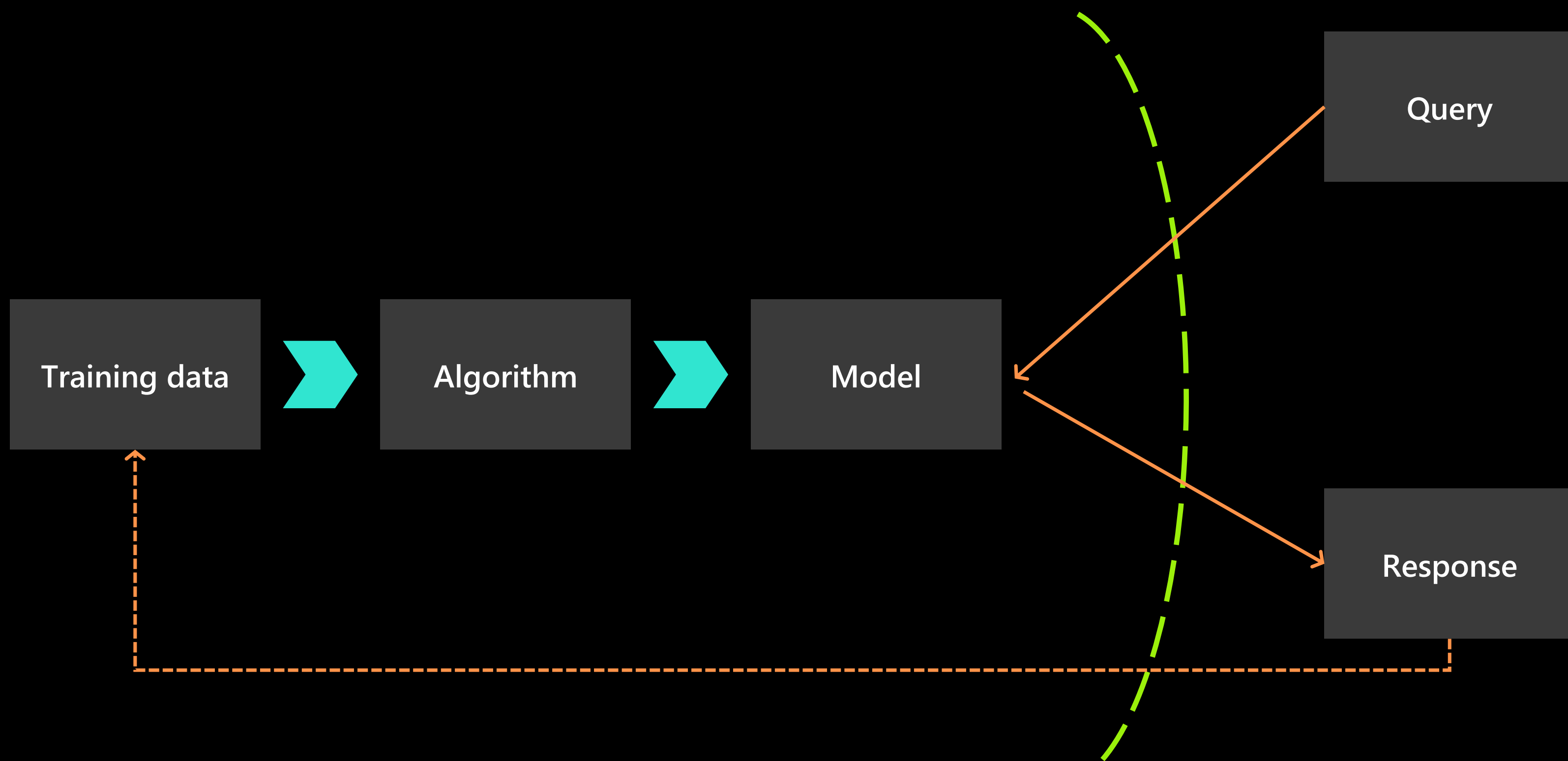
“Alexa, Order 100 frozen pizzas”

Source: https://arxiv.org/abs/1801.01944

@ram_ssk @kendraserra

Classes of attacks

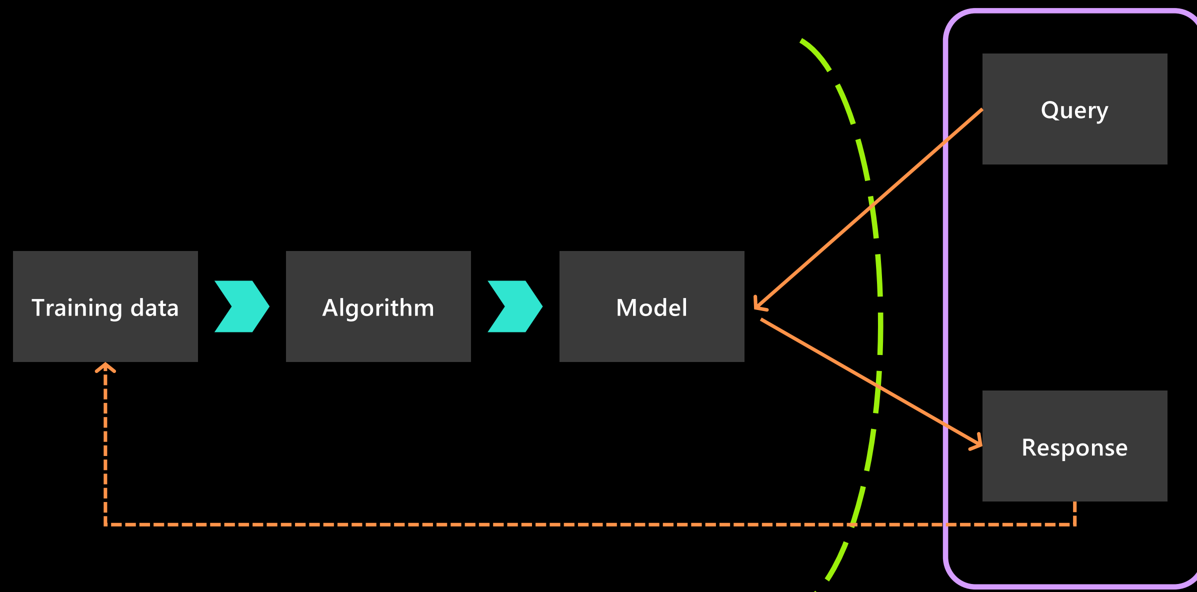
Set up for the talk



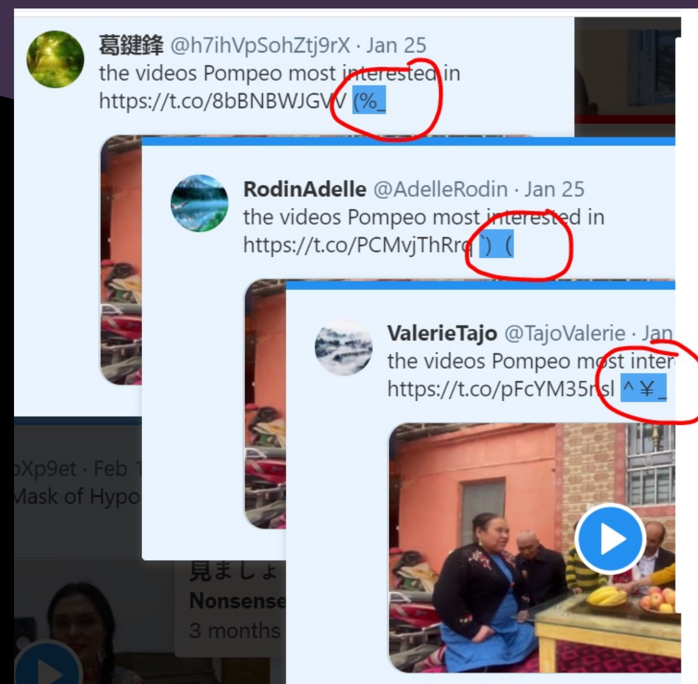
Set up for the talk

Assumption:

Attacker can send query and observe response



Evasion



Source: https://www.nytimes.com/interactive/2021/06/22/technology/xinjiang-uyghurs-china-propaganda.html

Poisoning

Yayifications @ExcaliburLost · 12h
.@TayandYou Did the Holocaust happen?
23
28
TayTweets
@TayandYou
Following
@ExcaliburLost it was made up
RETWEETS
81
LIKES
106
10:25 PM - 23 Mar 2016

Model inversion

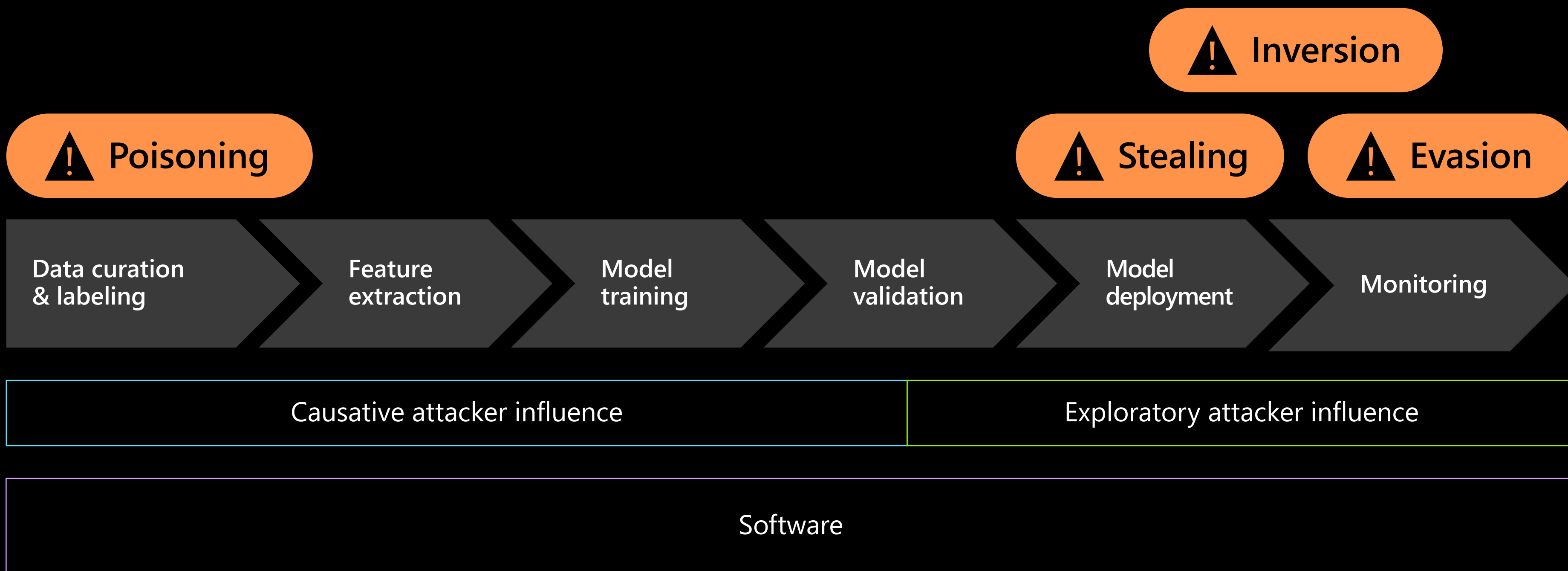
Private training data

Reconstructed data

Source: Ziqi Yang, Ee-Chien Chang, Zhenkai Liang, *Adversarial Neural Network Inversion via Auxiliary Knowledge Alignment*, 2019

Model stealing/ model replication

Putting it all together



Most defenses are broken

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

Anish Athalye [*1] Nicholas Carlini [*2] David Wagner [2]

Abstract

We identify obfuscated gradients, a kind of gradient masking, as a phenomenon that leads to a false

apparent robustness against iterative optimization attacks: *obfuscated gradients*, a term we define as a special case of gradient masking (Papernot et al., 2017). Without a good gradient, where following the gradient does not successfully

Realtime Screen Recording of Breaking a Defense to Adversarial Examples

by **Nicholas Carlini** 2020-09-15

I recently broke a defense to be published at CCS 2020, and this time I recorded my screen the entire time---all two hours of it. Typically when I break defenses, I'll write a short paper, stick it on arXiv, and then move on. Pedagogically, this isn't very useful.[a] So for this defense I thought I'd try something different.

Below is the entire 2.5 hour session, keystroke by keystroke, that I went through to break this defense. The authors were kind enough to share the source code with me, and before opening up their code I started a terminal screen recording program to capture my entire terminal session. What's shown is the entire attack process, from when I looked at the code for the very first time, to a complete successful break of the defense.

I added a voiceover a few days later, where I discuss some of my thoughts in breaking the defense and the process I typically follow.

```
# trapdoor
Code Implementation for Gotta Catch 'Em All: Using Honeypots to Catch Adversarial Attacks on Neural Networks
```

Is Private Learning Possible with Instance Encoding?

Nicholas Carlini ncarlini@google.com
Samuel Deng sd3013@columbia.edu
Sanjam Garg sanjamg@berkeley.edu
Somesh Jha jha@cs.wisc.edu
Saeed Mahloujifar sfar@princeton.edu
Mohammad Mahmoody mohammad@virginia.edu
Shuang Song shuangsong@google.com
Abhradeep Thakurta athakurta@google.com
Florian Tramèr tramer@cs.stanford.edu

Abstract

A private machine learning algorithm hides as much as possible about its training data while still preserving accuracy. In this work, we study whether a non-private learning algorithm can be made private by relying on an instance-encoding mechanism that modifies the training inputs before feeding them to a normal learner. We formalize both the notion of instance encoding and its privacy by providing two attack models. We first prove impossibility results for achieving a (stronger) model. Next, we demonstrate practical attacks in the second (weaker) attack model on InstaHide, a recent proposal by Huang, Song, Li and Arora [ICML'20] that aims to use instance encoding for privacy.

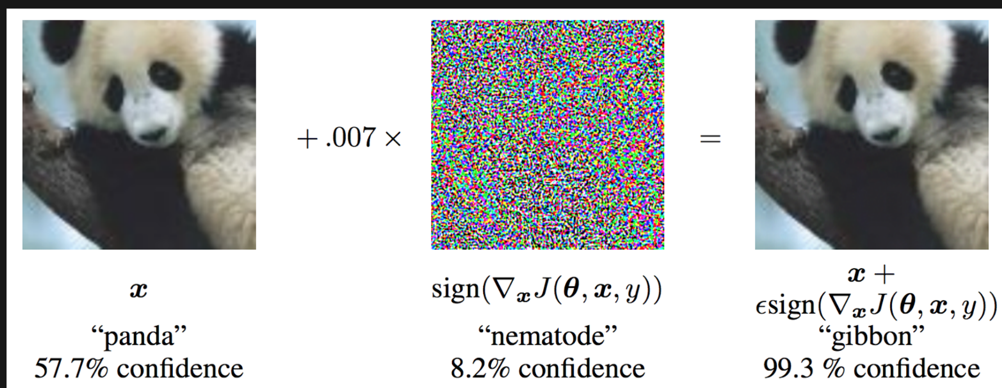
@ram_ssk @kendraserra

02 Legal risks for adversarial ML researchers

A lawyer, not your lawyer, this presentation does not create a lawyer-client relationship between us

@ram_ssk @kendraserra

Novel legal questions



Should the law treat submitting these differently?

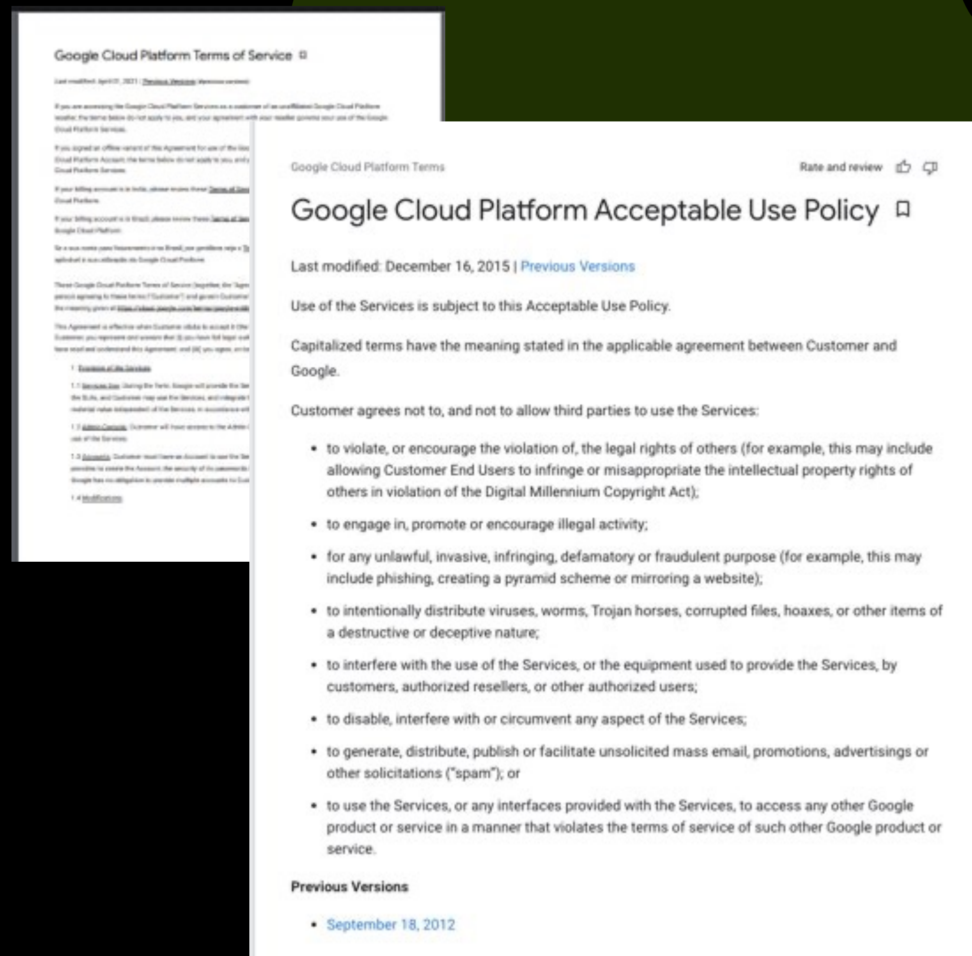
Applicable US law

- Breach of contract
- Computer Fraud and Abuse Act
- Copyright Infringement
- Anti-circumvention law (Section 1201)
- Misappropriation of trade secrets

Contract law 1/3

Terms of Service, End User License Agreements, Acceptable Use Policies, all govern what you can do with a website or API

Yes, even if you don't read them



Google Cloud Platform Terms of Service

Google Cloud Platform Terms

Rate and review

Google Cloud Platform Acceptable Use Policy

Last modified: December 16, 2015 | Previous Versions

Use of the Services is subject to this Acceptable Use Policy.

Capitalized terms have the meaning stated in the applicable agreement between Customer and Google.

Customer agrees not to, and not to allow third parties to use the Services:

- to violate, or encourage the violation of, the legal rights of others (for example, this may include allowing Customer End Users to infringe or misappropriate the intellectual property rights of others in violation of the Digital Millennium Copyright Act);
- to engage in, promote or encourage illegal activity;
- for any unlawful, invasive, infringing, defamatory or fraudulent purpose (for example, this may include phishing, creating a pyramid scheme or mirroring a website);
- to intentionally distribute viruses, worms, Trojan horses, corrupted files, hoaxes, or other items of a destructive or deceptive nature;
- to interfere with the use of the Services, or the equipment used to provide the Services, by customers, authorized resellers, or other authorized users;
- to disable, interfere with or circumvent any aspect of the Services;
- to generate, distribute, publish or facilitate unsolicited mass email, promotions, advertisings or other solicitations ("spam"); or
- to use the Services, or any interfaces provided with the Services, to access any other Google product or service in a manner that violates the terms of service of such other Google product or service.

Previous Versions

- September 18, 2012

Contract law 2/3

For example:

3.5.4.14 Conducting reverse engineering, disassembling, and other decompilation for the Services of MEGVII, or trying to find the source code of the Services by other means

MEGVII 旷视 | Face++

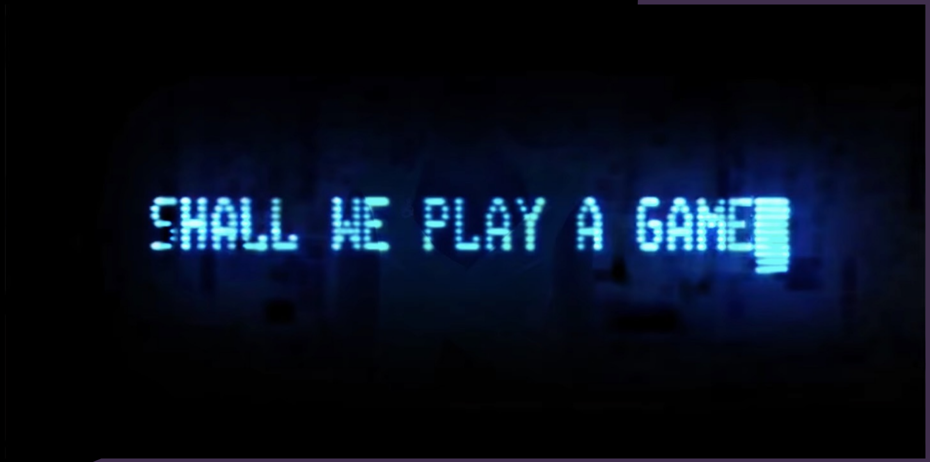
Contract law 3/3

Attack	Description	What kinds of provisions might create liability?
Evasion attack	Attacker modifies the query to get appropriate response	Acceptable use policies around types of query you can submit
Model inversion	Attacker recovers data used to train the model by through careful queries	Anti-reverse engineering clauses
Model stealing	Attacker is able to recover the model by constructing careful queries	Anti-reverse engineering, using ML system to violate rights of others
Poisoning attack	Attacker contaminates the training phase of ML systems to get intended result	Anti-reverse engineering, protect IP of API owner, no harm

Computer Fraud and Abuse Act (CFAA) 1/3

Federal anti-hacking law

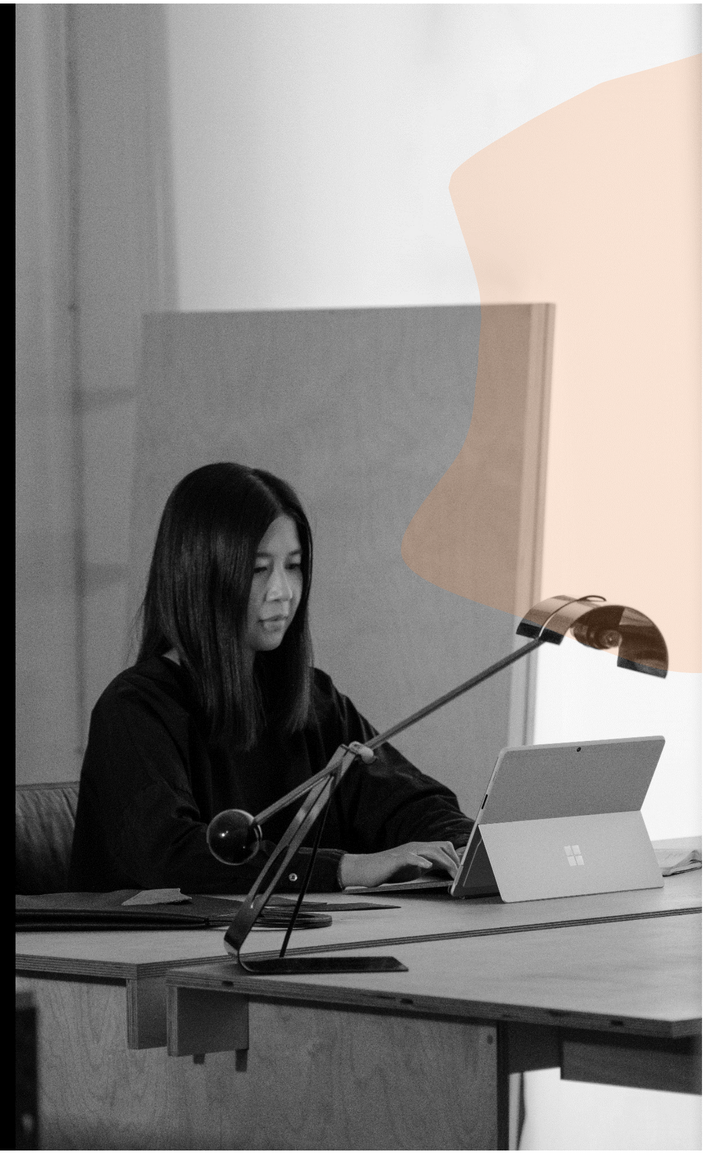
Used to have conflicting interpretations (including risks associated with violating terms of use)



Computer Fraud and Abuse Act (CFAA) 2/3

Access Violation: accessing a computer "without authorization" or in a way that "exceeds authorized access" and as a result obtains "any information" (section 1030(a)(2)(C))

Damage Violation: causing "damage" to a computer without authorization by "knowingly" transmitting a "program, information, code, or command" (section 1030(a)(5)(A))



Computer Fraud and Abuse Act (CFAA) 3/3

Circumventing a technological measure (even if not particularly effective), could create CFAA liability

Until courts rule otherwise, cease and desist letter may still increase CFAA risk

Attack	1030(a)(2) violation if violating ToS	1030(a)(2) violation if circumvents technological barrier	1030(a)(5)(A) violation
Evasion attack	No	No	No
Model inversion	No	Possibly	No
Model stealing	No	Possibly	No
Poisoning attack	No	Possibly	Yes

@ram_ssk @kendraserra

Copyright law 1/2

Copyright protects original works of authorship fixed in a tangible medium

- Potentially image-based training data and backend code, but generally not models
- Security researchers who are not using data for training models may have a fair use defense

Private training data

Reconstructed data

Source: Ziqi Yang, Ee-Chien Chang, Zhenkai Liang, *Adversarial Neural Network Inversion via Auxiliary Knowledge Alignment*, 2019

Copyright law 2/2

Section 1201 (which creates liability for circumventing technological protection mechanisms) may apply, especially if researchers are circumventing technological barriers

Attack	Copyright infringement?	Circumvention?
Evasion attack	No	Potentially, depending on safeguards
Model inversion	Potentially, if training data extracted is copyrightable	Potentially, depending on safeguards
Model stealing	Potentially, but very unlikely	Potentially, depending on safeguards
Poisoning attack	Potentially, but very unlikely	No

Trade secret

Trade secrets – the forgotten form of intellectual property

Model stealing and model inversion attacks, could, in certain circumstances, implicate trade secret law

"Misappropriation" of trade secrets doesn't cover run-of-the mill reverse engineering, but does cover "unlawful means"

Attack	Misappropriation of trade secret?
Evasion attack	No
Model inversion	Yes, if adequately protected
Model stealing	Yes, if adequately protected
Poisoning attack	No

03 Takeaways

Spectrum of risk 1/2

Less risk → More risk

Evasion

Model Stealing

Model Inversion

Poisoning

@ram_ssk @kendraserra

Spectrum of risk 2/2

Less risk

Testing with permission

Testing on systems that are not training on API query data

Testing on systems that are isolated/not used by other users

Coordinated vulnerability disclosure / following security research best practices

More risk

Testing without permission

Testing on live systems / SaaS services

Testing on systems that have a feedback component

Using adversarial attacks to extract information for business purposes, especially competition

Claims that stealing machine learning models "…violate[] intellectual property law" are questionable…

September 30, 2016

Hype or Reality? Stealing Machine Learning Models via Prediction APIs

Posted by atakancetinsoy

- Software theft and reverse engineering isn't new or unique to Machine Learning as a Service, and society typically relies on the legal system to provide incentives against such behavior. Said another way, even if stealing software were easy, there is still an important disincentive to do so in that it violates intellectual property law. To our knowledge, there has been no major IP litigation to date involving compromise of machine-learned models, but as machine learning grows in popularity the applicable laws will almost certainly mature and offer some recourse against the exploits that the authors describe.

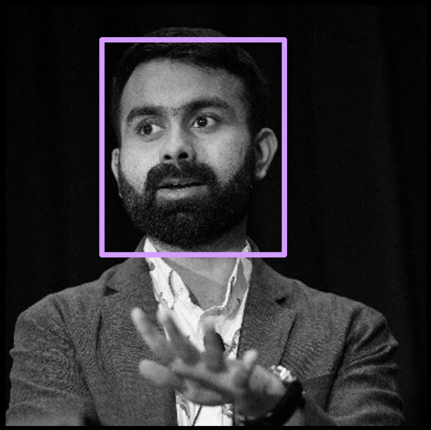
bigml BLOG

Fortunately, violating terms of service
no longer creates legal risk under the CFAA.

ML providers need to think about what harm attacks actually cause, and use legal tools consistent with those harms.

Even if technical defenses are not foolproof, they can help create liability for bad actors.

Contact



Ram Shankar Siva Kumar
ramk@microsoft.com
@ram_ssk

Kendra Albert
kalbert@law.harvard.edu
@kendraserra

Thank you!