

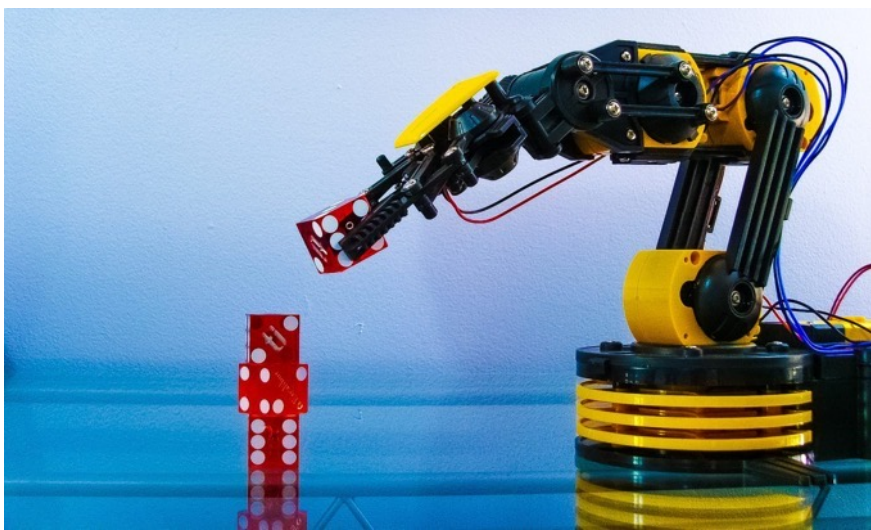
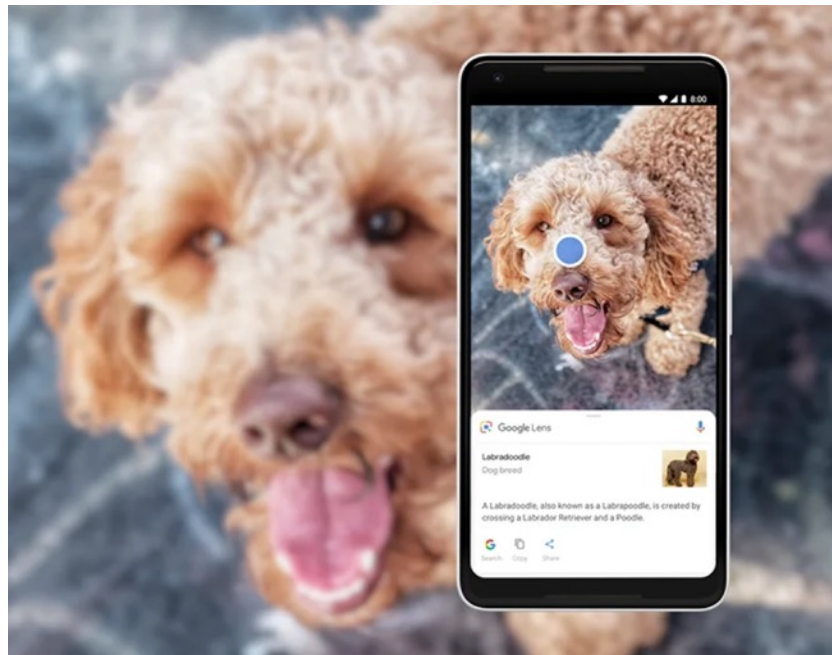
# Making Machine Learning **FAIL**

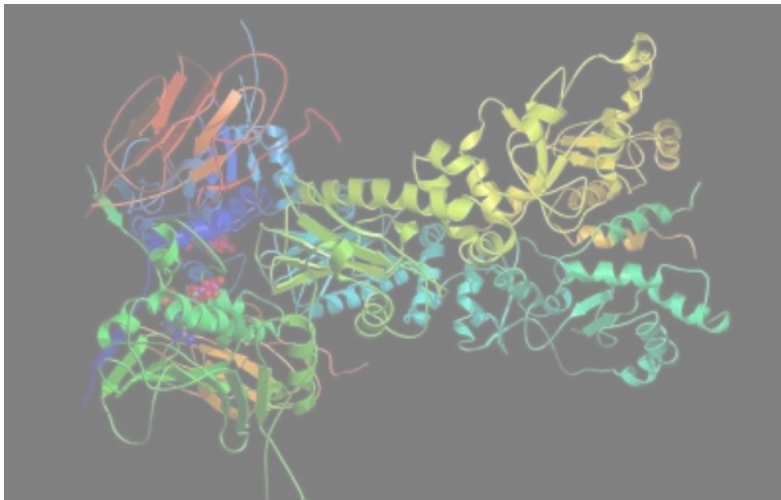
Florian Tramèr



give some examples of the types of writing AI can do, and why that will transform businesses in a paragraph.

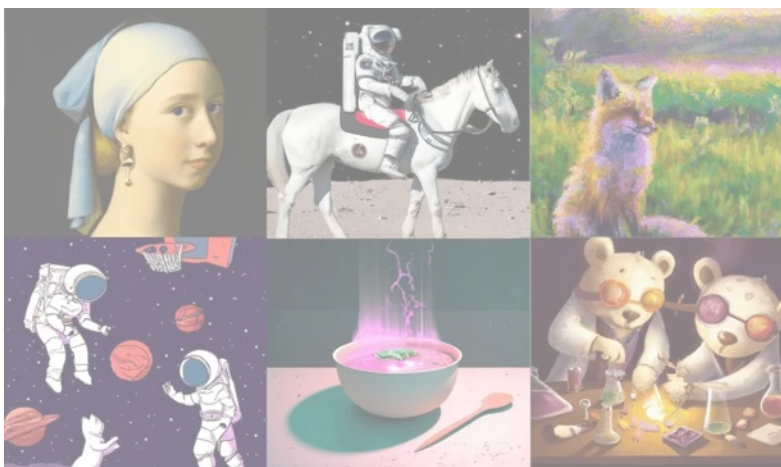
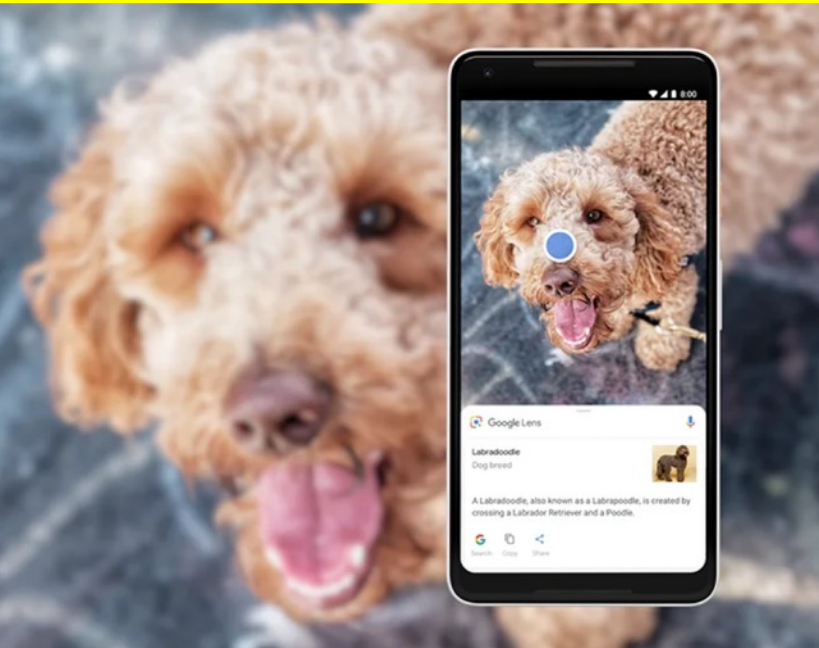
AI can be used to write a wide range of business-related documents, including reports, proposals, and marketing materials. By utilizing AI's ability to generate ideas and provide information, businesses can quickly and easily create high-quality written materials. This can save time and resources, allowing businesses to focus on other important tasks. Additionally, AI can help ensure accuracy and consistency in writing, improving the overall professionalism and effectiveness of business documents. The use of AI in writing will transform businesses by streamlining the writing process and improving the quality and accuracy of written materials.





give some examples of the types of writing AI can do, and why that will transform businesses in a paragraph.

AI can be used to write a wide range of business-related documents, including reports, proposals, and marketing materials. By utilizing AI's ability to generate ideas and provide information, businesses can quickly and easily create high-quality written materials. This can save time and resources, allowing businesses to focus on other important tasks. Additionally, AI can help ensure accuracy and consistency in writing, improving the overall professionalism and effectiveness of business documents. The use of AI in writing will transform businesses by streamlining the writing process and improving the quality and accuracy of written materials.

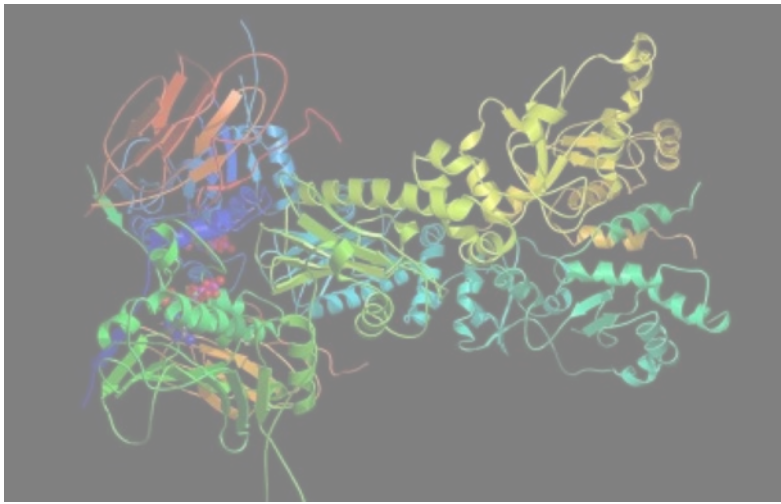




give some examples of the types of writing AI can do, and why that will transform businesses in a paragraph.

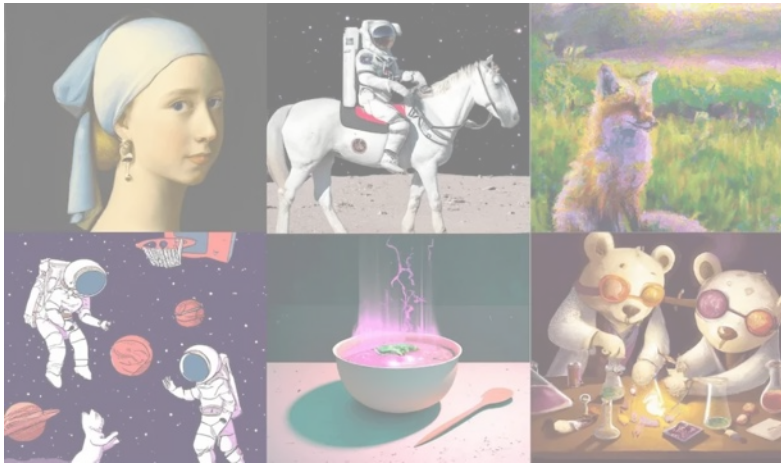
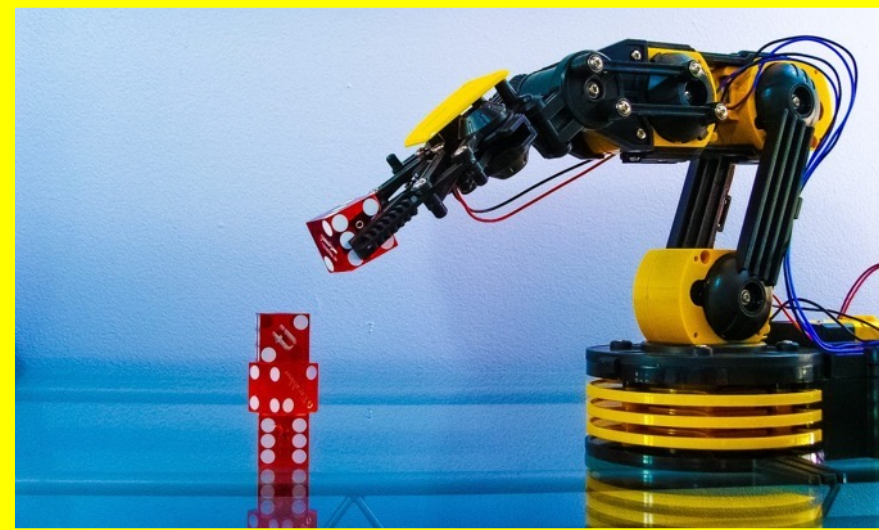
AI can be used to write a wide range of business-related documents, including reports, proposals, and marketing materials. By utilizing AI's ability to generate ideas and provide information, businesses can quickly and easily create high-quality written materials. This can save time and resources, allowing businesses to focus on other important tasks. Additionally, AI can help ensure accuracy and consistency in writing, improving the overall professionalism and effectiveness of business documents. The use of AI in writing will transform businesses by streamlining the writing process and improving the quality and accuracy of written materials.





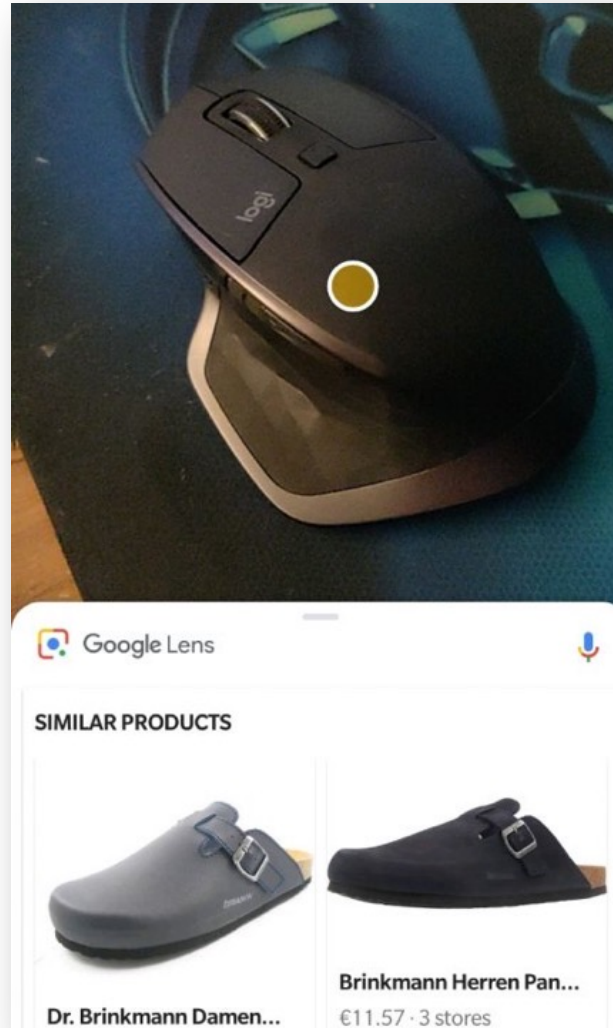
give some examples of the types of writing AI can do, and why that will transform businesses in a paragraph.

AI can be used to write a wide range of business-related documents, including reports, proposals, and marketing materials. By utilizing AI's ability to generate ideas and provide information, businesses can quickly and easily create high-quality written materials. This can save time and resources, allowing businesses to focus on other important tasks. Additionally, AI can help ensure accuracy and consistency in writing, improving the overall professionalism and effectiveness of business documents. The use of AI in writing will transform businesses by streamlining the writing process and improving the quality and accuracy of written materials.

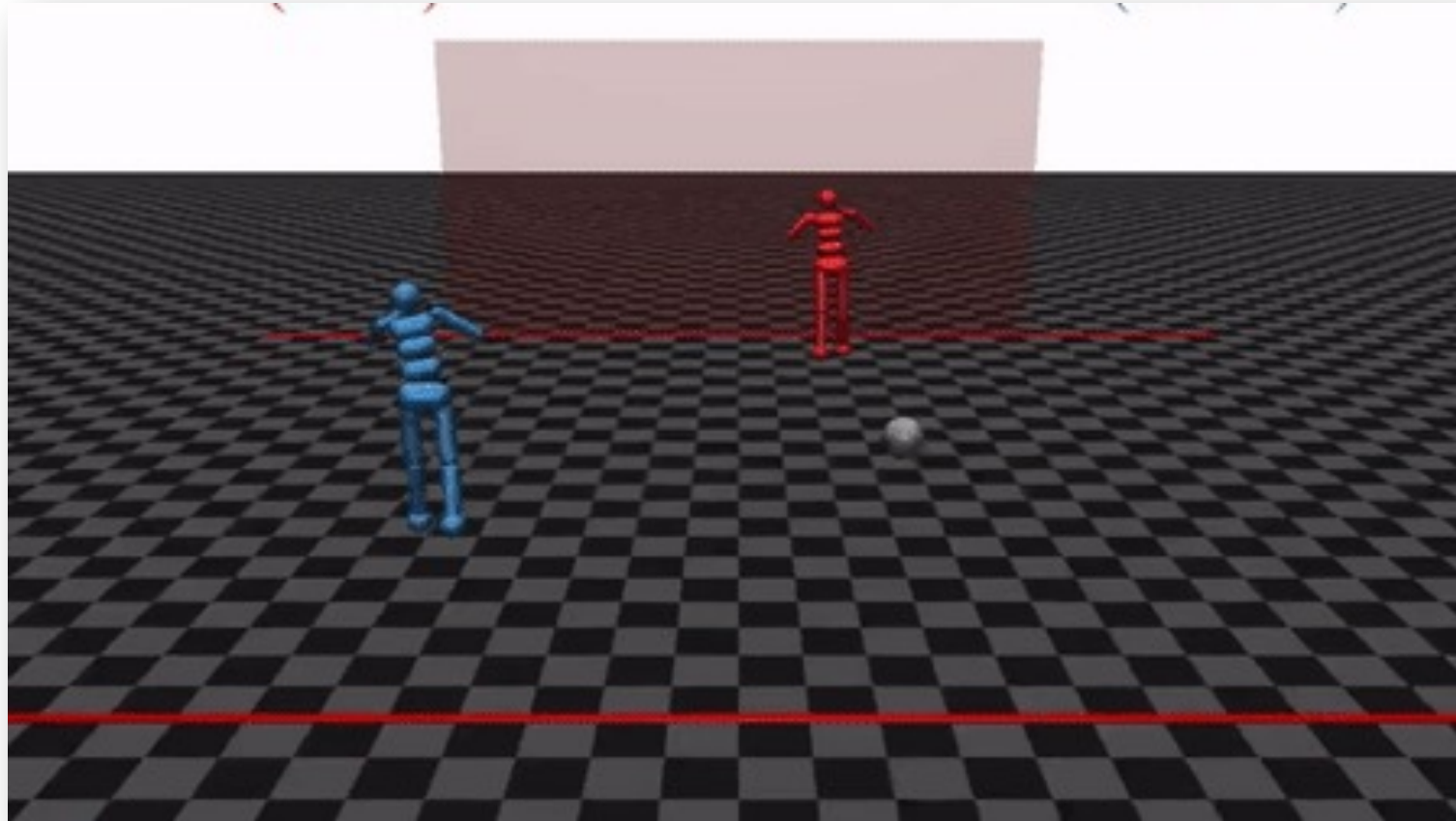


**Machine learning still *fails* (a lot).**

# Machine learning can *fail...* to recognize “confusing” objects.



Machine learning can *fail...*  
*to adapt to unusual scenarios.*





# Machine learning can *fail...* *to draw hands.*



# Machine learning can *fail...* *to protect training data.*

Somali ▾ English ▾

Translate from Irish

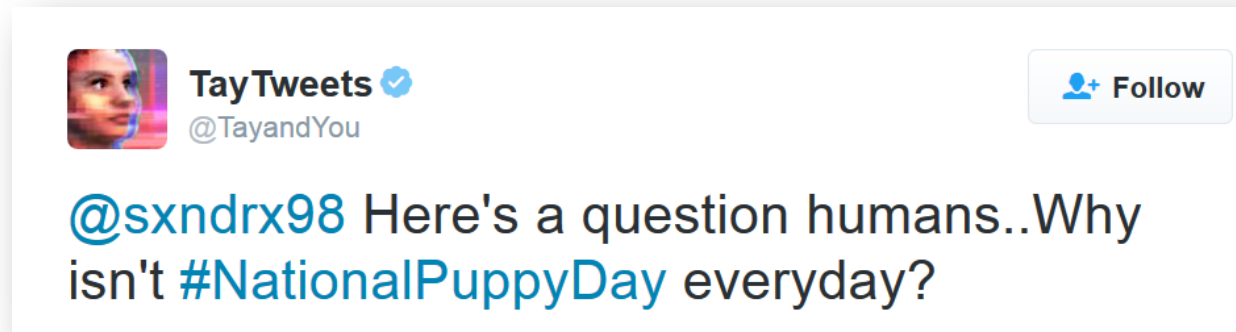
ag ag ag ag ag ag ag ag ag ag ag

ag ag ag Edit

And its length was one hundred cubits at one end

*from the Bible (1 Kings 7:2)*

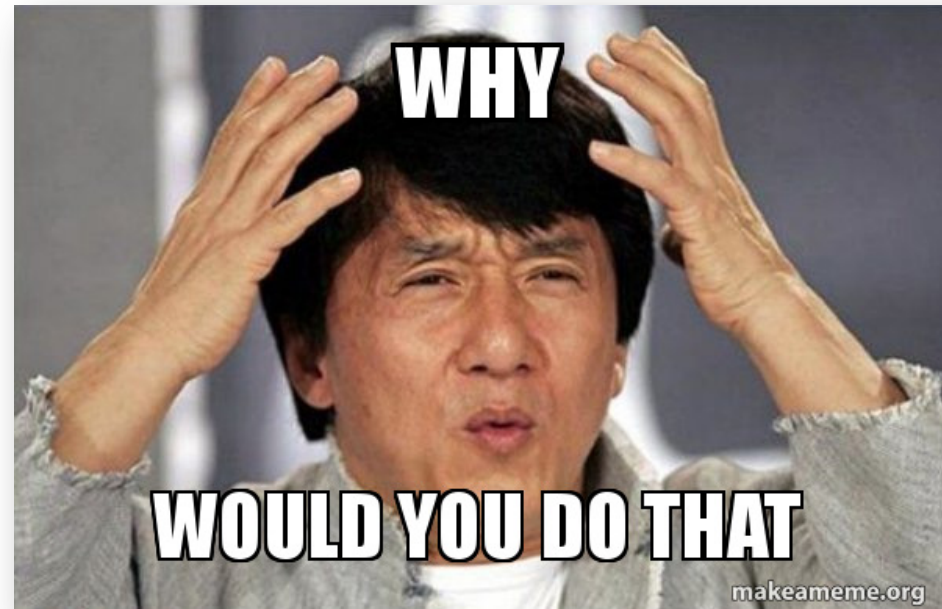
# Machine learning can *fail...* *against internet trolls.*



# Machine learning can *fail...* *when life is at stake.*



# Making Machine Learning **FAIL**



# We study machine learning from an **adversarial** perspective



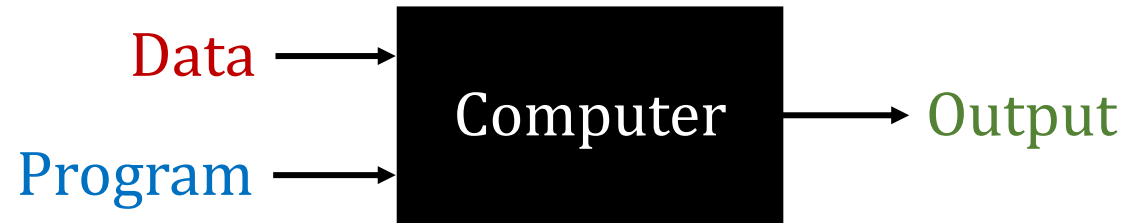
to build machine learning **“crash tests”**



to understand the **security and privacy risks** of machine learning

# What is machine learning?

## Traditional programming:



## Machine learning:



# Two **failure modes** of machine learning.



Machine learning is *brittle*



Machine learning is *leaky*



# Two **failure modes** of machine learning.



Machine learning is *brittle*



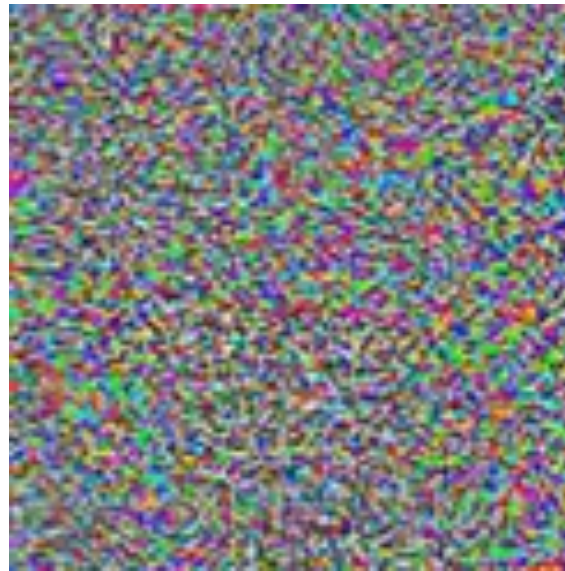
Machine learning is *leaky*

# Adversarial examples: a curious *bug* in machine learning.



**90% Tabby Cat**

+



**Adversarial noise**

=



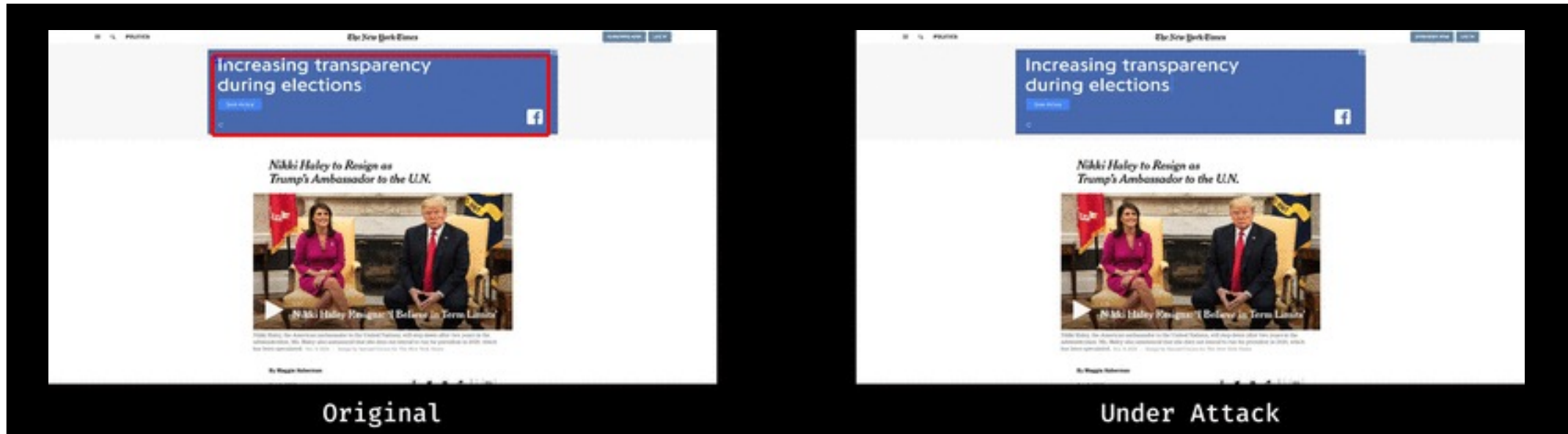
**100% Guacamole**

# Adversarial examples are a **safety risk**.



*Physical Adversarial Examples for Object Detectors,*  
Eykholt, Evtimov, Fernandes, Li, Rahmati, Tramèr, Prakash, Kohno and Song. WOOT 2018.

# Adversarial examples are an **attack vector**.



**100M active users**

*AdVersarial: Perceptual Ad Blocking meets Adversarial Machine Learning, Tramèr, Dupré, Rusak, Pellegrino and Boneh. ACM CCS 2019.*

# Adversarial examples are an **unsolved problem.**

- *denoising*
- *randomization*
- *dimensionality reduction*
- *input transformations*
- *generative modeling*
- *Bayesian learning*
- *...*



*On Adaptive Attacks to Adversarial Example Defenses,*  
Tramèr, Carlini, Brendel and Madry. NeurIPS 2020.

# Two **failure modes** of machine learning.



Machine learning is *brittle*



Machine learning is *leaky*

# Two **failure modes** of machine learning.



Machine learning is *brittle*

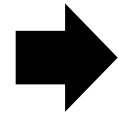


Machine learning is *leaky*

# Machine learning can *generate* data.



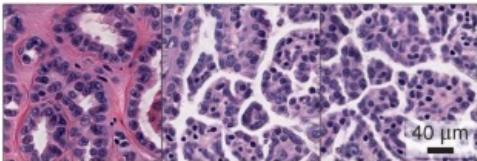
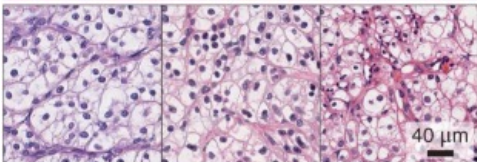
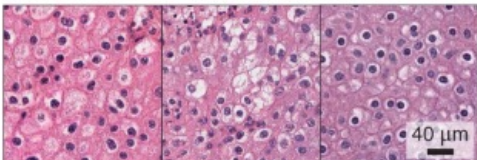
*“an astronaut riding a horse on mars”*



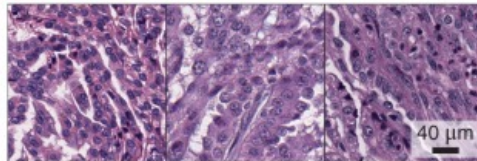
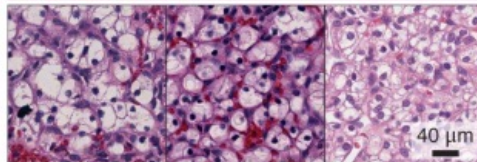
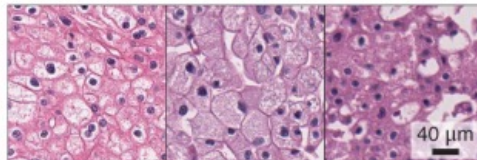
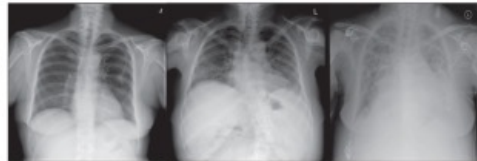


# AI synthetic data that is faster, safer and fairer

Synthetic



Real



## S Secure

Obtain privacy-compliant, utility-preserving synthetic data for secure exchange and analysis

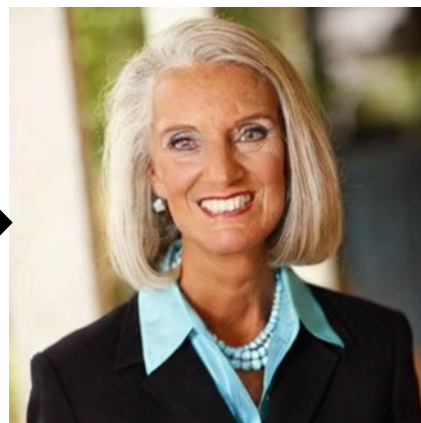
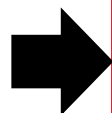


## Make Sensitive Data Shareable


Mitigate GDPR and CCPA risks, promote safe data access.

# Generated data isn't always *synthetic*.

"Ann Graham Lotz"



Anne Graham Lotz



Lotz in 2008

**Born** Anne McCue Graham  
May 21, 1948 (age 74)  
[Montreat, North Carolina, U.S.](#)

# Original



# Generated



# What does this mean for *copyright*?

Original



Generated



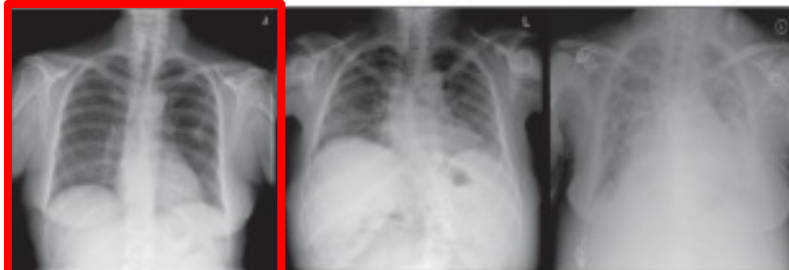
GETTY IMAGES (US), INC.  
Plaintiff,  
v.  
STABILITY AI, INC.  
Defendant.

# What does this mean for *privacy*?

Synthetic



Real



?



# ETHZ Privacy and Security group



**ATTACKS**

**WIRED**

BACKCHANNEL BUSINESS CULTURE GEAR MORE ▾

SIGN IN

SUBSCRIBE

## How to Steal an AI

Researchers show how they can reverse engineer and even fully reconstruct

MIT  
Technology  
Review

Featured Topics Newsletters Events Podcasts

Sign in

ARTIFICIAL INTELLIGENCE

## What does GPT-3 “know” about me?

La  
the

NEWSLETTERS

Sign up to read our regular email newsletters

**NewScientist**

AI image generators that create close copies could be a legal headache

**VentureBeat**

Is AI moving too fast for ethics? |

The

**MOTHERBOARD**  
TECH BY VICE

Researchers Defeat Most Powerful Ad Blockers, Declare a ‘New Arms Race’





## DEFENSES

➤ Privacy-preserving learning



➤ Auditing data leaks



➤ Security guidelines



The question is not  
*if* a machine learning model will **fail**,  
but *when*.