Lecture 23:
Ethics in Machine Learning

Slides adapted from Alina Vereshchaka

GOAL OF AI

What is the Goal of AI?

OECD's* Principles on AI:

"AI should benefit people and the planet"

*__Organization for Economic Co-operation and Development (OECD)__ is an intergovernmental economic organization with 36 member countries to stimulate economic progress and world trade.

Source: https://www.oecd.org/going-digital/ai/principles/

Ethics in AI

Humans are intelligent to the extent that **our** actions can be expected to achieve **our** objectives.

Machines are intelligent to the extent that **their** actions can be expected to achieve **their** objectives

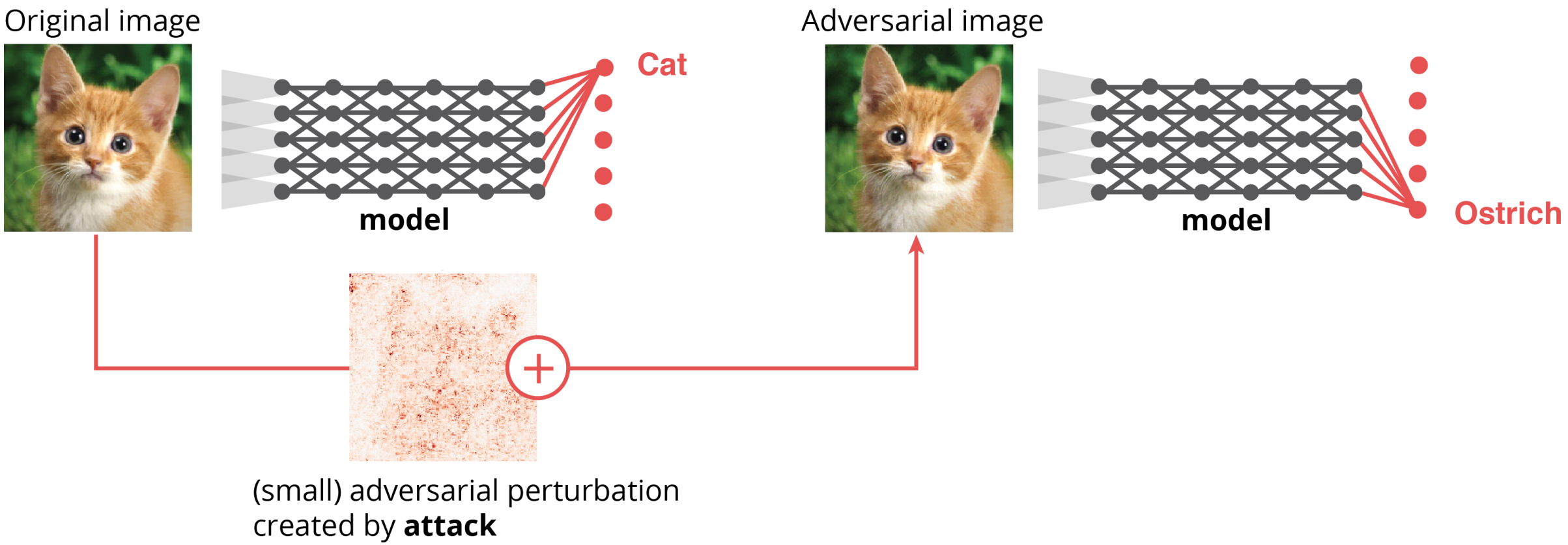
- Control theory: minimize cost function
- Economics: maximize expected utility
- Operations research: maximize sum of rewards
- Statistics: minimize loss function
- AI: all of the above, plus logically defined goals

Machines are **beneficial** to the extent that **their** actions can be expected to achieve **our** objectives

We need machines to be **provably beneficial**


ADVERSARIAL ATTACKS

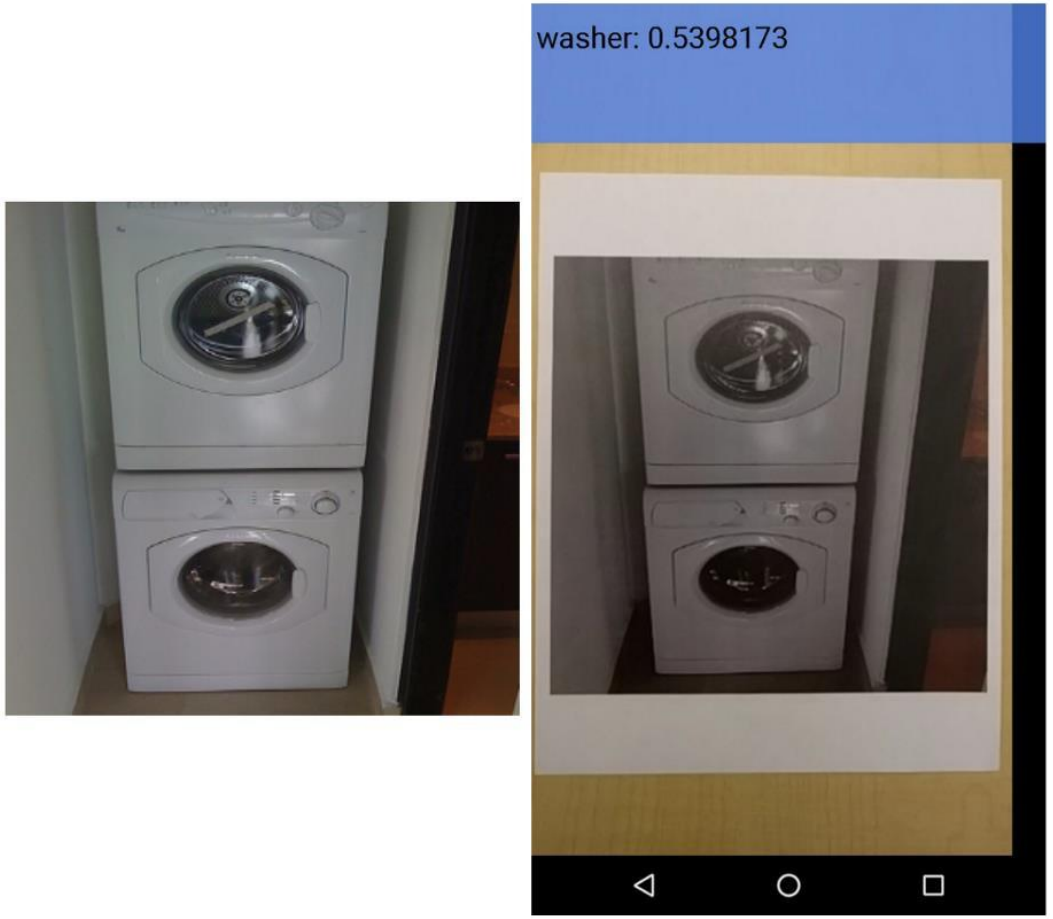
Adversarial Attacks



Adversarial Attacks

Source: [Adversarial Examples in the Physical World](). Kurakin et al, ICLR 2017.

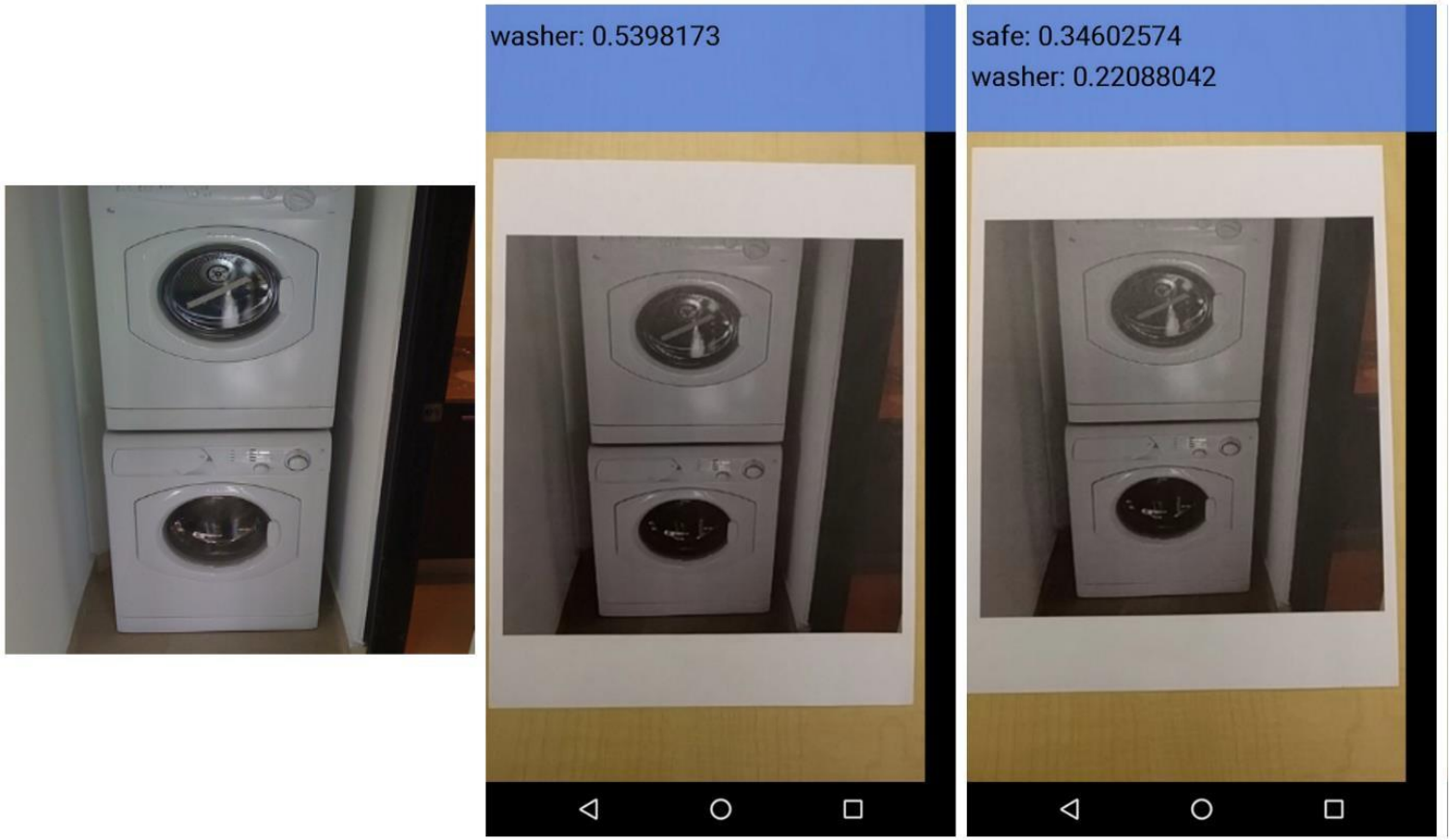
Adversarial Attacks



(a) Image from dataset (b) Clean image

Source: [Adversarial Examples in the Physical World](). Kurakin et al, ICLR 2017.

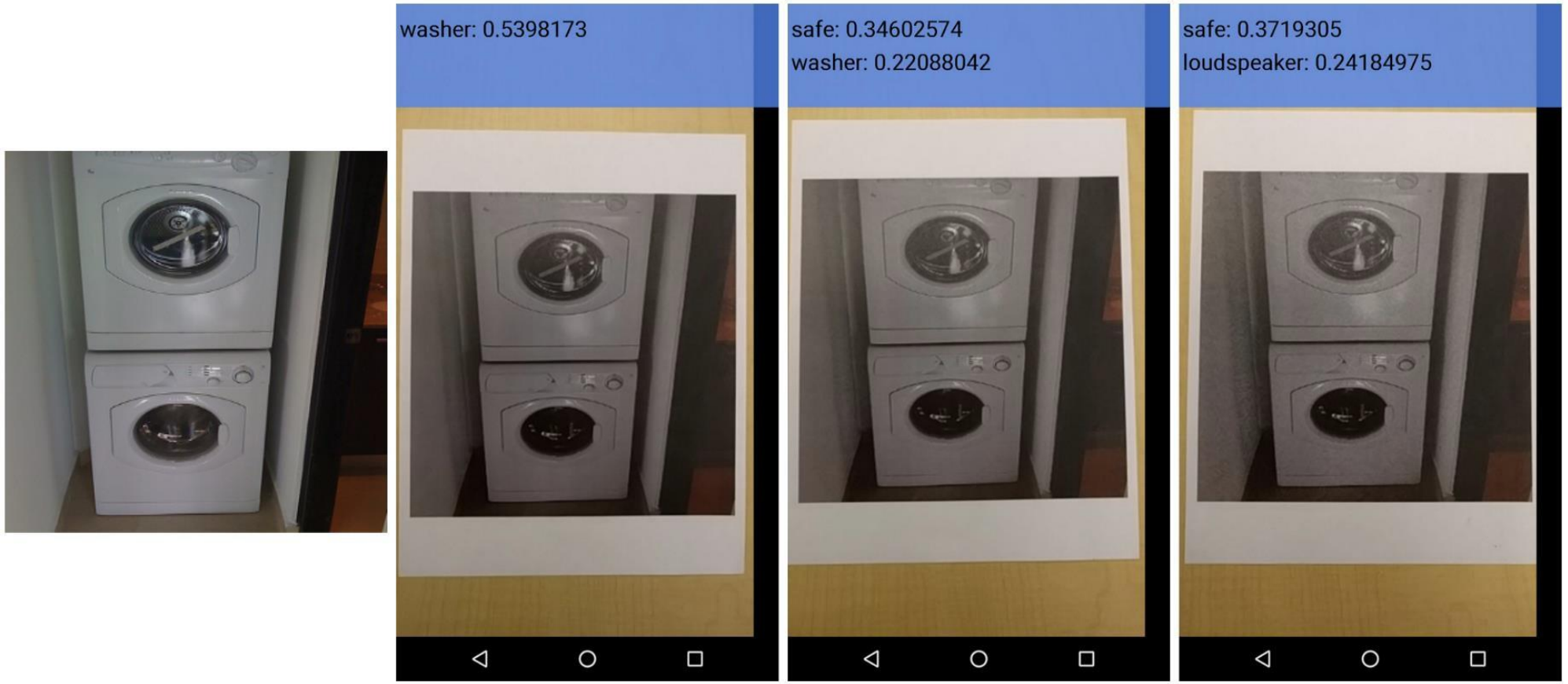
Adversarial Attacks



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$

Source: [Adversarial Examples in the Physical World](). Kurakin et al, ICLR 2017.

Adversarial Attacks



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

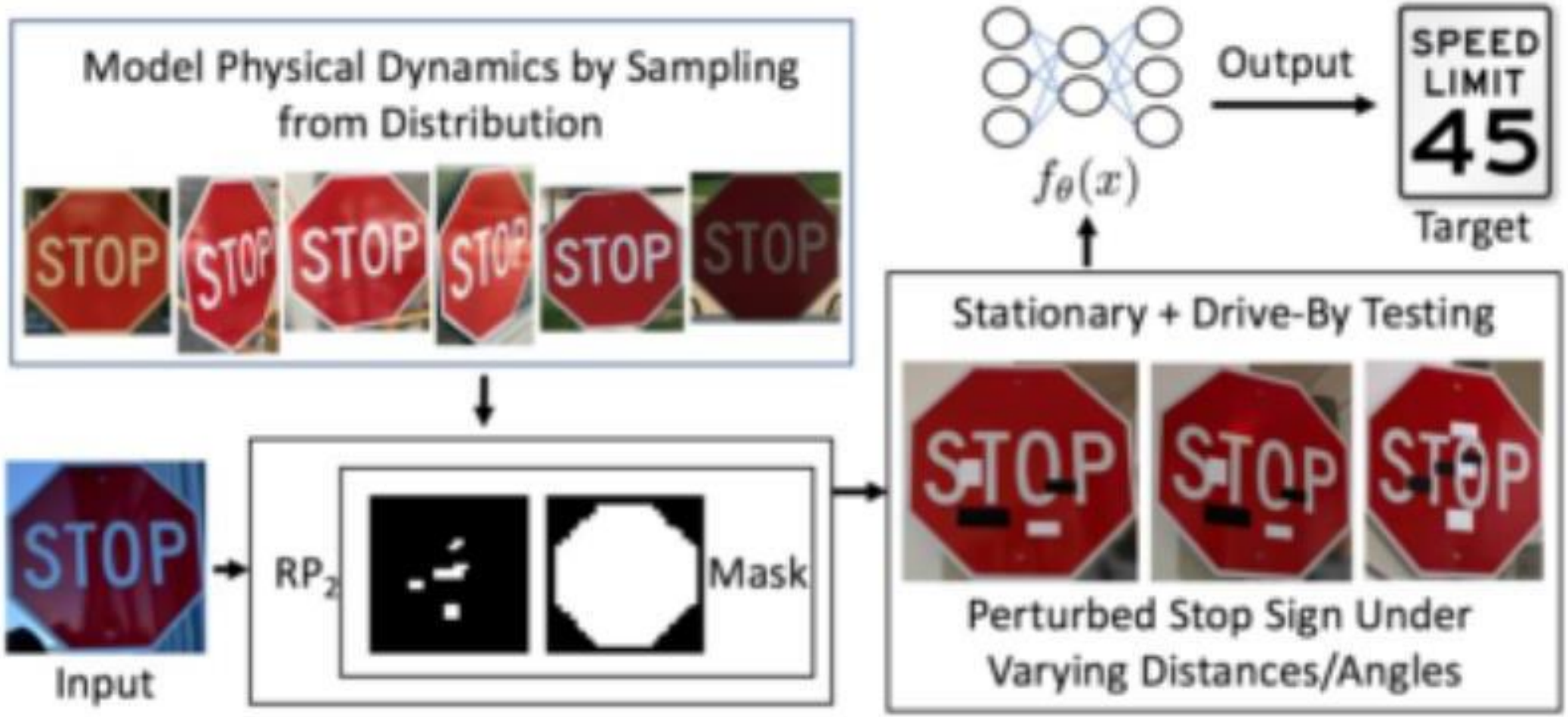
Source: [Adversarial Examples in the Physical World](). Kurakin et al, ICLR 2017.

Adversarial Attacks

Figure 1: The left image shows real graffiti on a Stop sign, something that most humans would not think is suspicious. The right image shows our a physical perturbation applied to a Stop sign. We design our perturbations to mimic graffiti, and thus "hide in the human psyche."

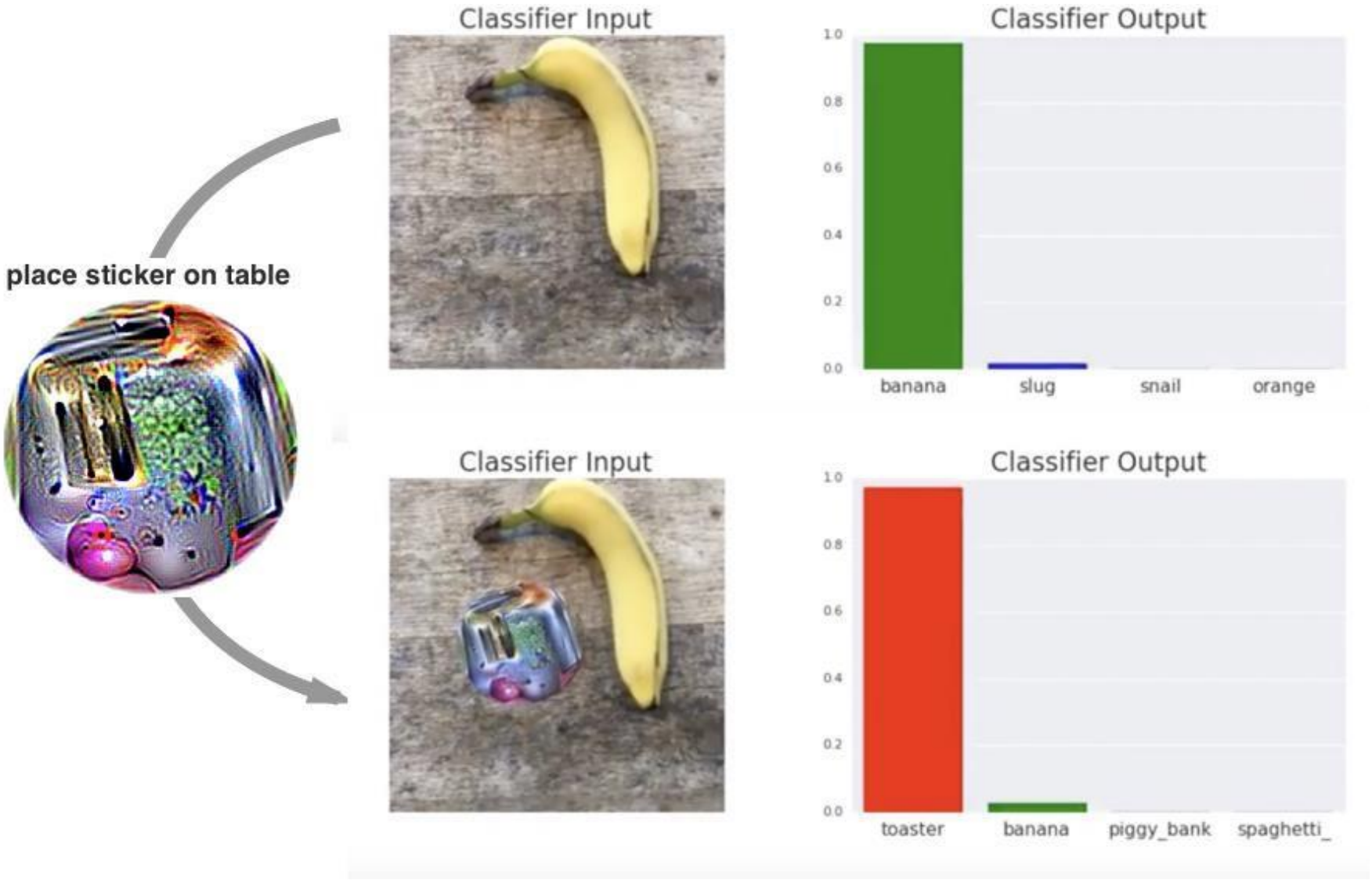
Source: Robust Physical-World Attacks on Deep Learning Visual Classification

Adversarial Attacks



Source: Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR 2018

Adversarial Attacks



Source: Adversarial Patch: https://arxiv.org/pdf/1712.09665.pdf

How can we prevent adversarial attacks?

Possible solutions

- **ADVERSARIAL TRAINING**

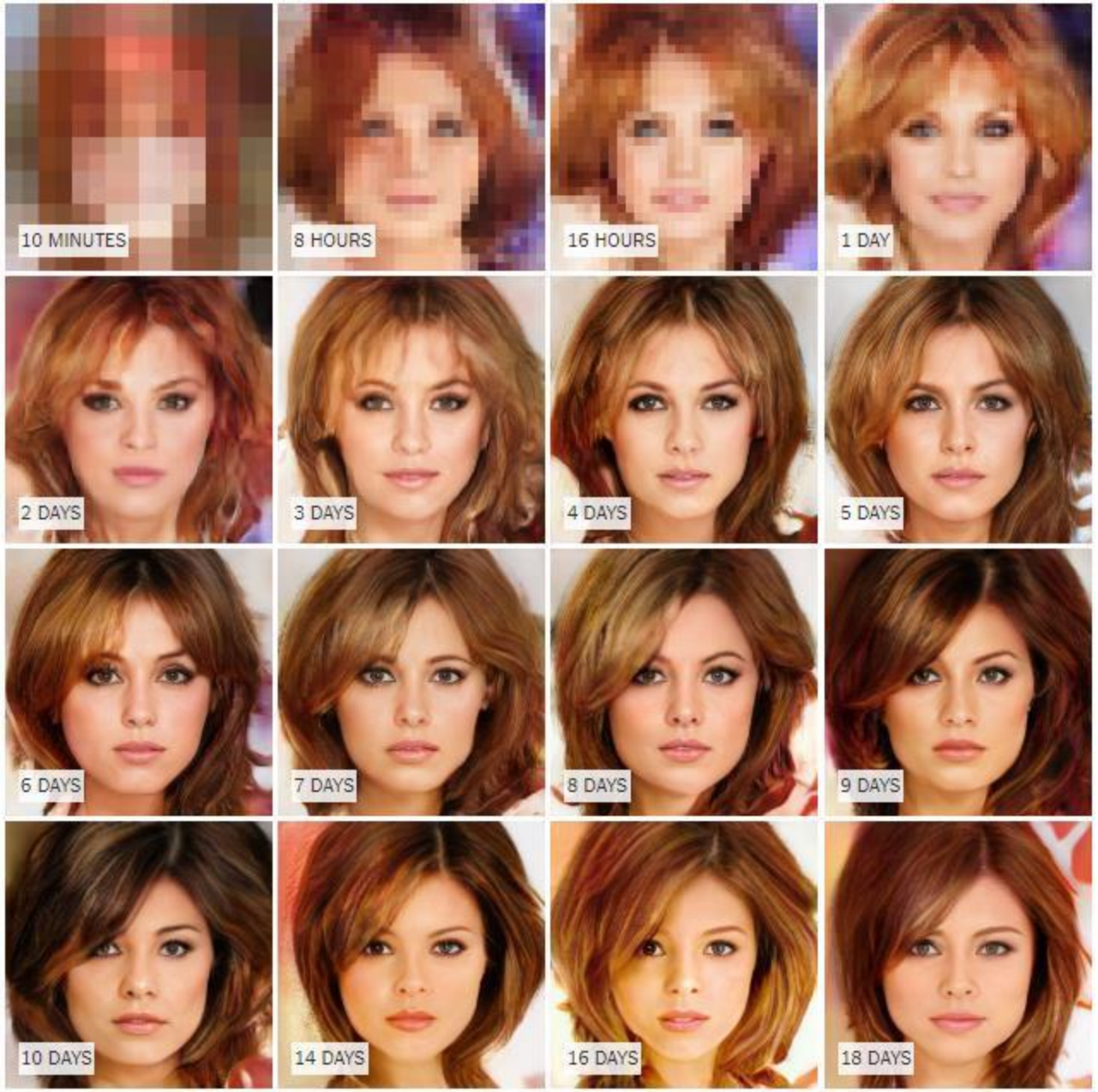
Pretend to be the attacker, generate a number of adversarial examples against your own network, and then explicitly train the model to not be fooled by them.

- **DEFENSIVE DISTILLATION**

Train a secondary model whose surface is smoothed in the directions an attacker will typically try to exploit, making it difficult for them to discover adversarial input tweaks that lead to incorrect categorization.

FAKE NEWS & FAKE VIDEOS

DeepFaces



This series of images shows the output of Nvidia's system over the course of 18 days of processing. With their method, called progressive GANs, the Nvidia researchers built a system that begins with low-resolution images and then gradually progresses to higher resolutions. This allows the training to happen more quickly, but it also in a more controlled and stable way. The result: 1024- by 1024-pixel images that are sharp, detailed, and, in many cases, very convincing. Source: Nvidia

https://thispersondoesnotexist.com/

Generated Videos

OpenAI Sora examples

Deepfake example

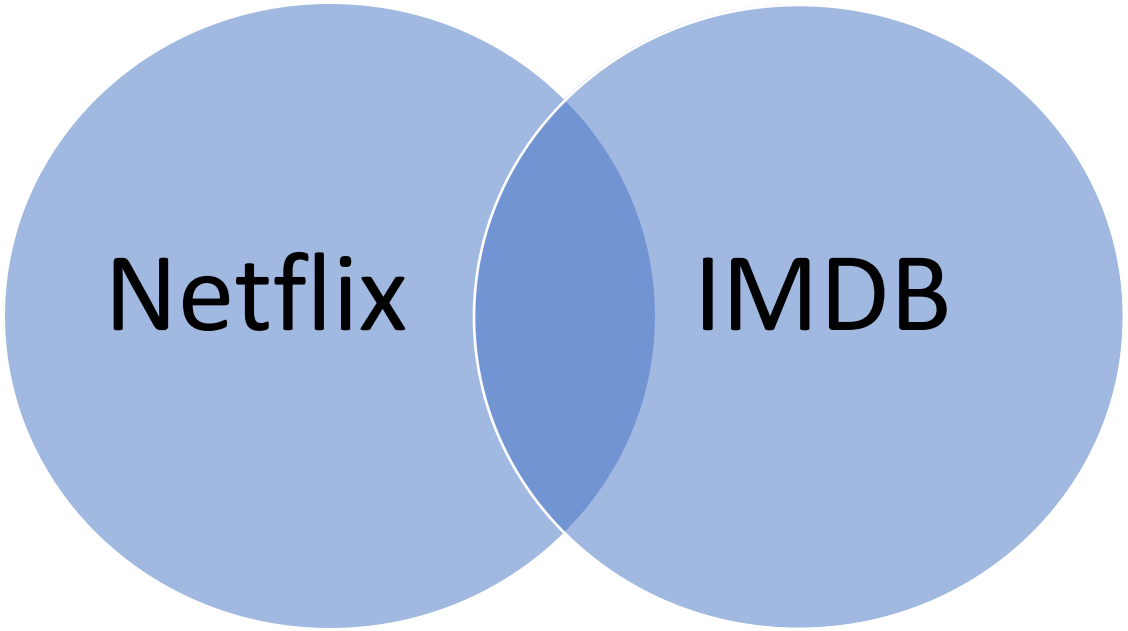
If we know that videos can be faked, what will be acceptable as evidence?

How can we differentiate between real/generated data?

Combating Misinformation

DIFFERENTIAL PRIVACY

Why Anonymization is hard?



Netflix Challenge (2006), a Kaggle-style competition to improve their movie recommendations, with a $1 million prize

- They released a dataset consisting of 100 million movie ratings (by "anonymized" numeric user ID), with dates
- Researchers found they could identify 99% of users who rated 6 or more movies by cross-referencing with IMDB, where

Why is anonymization hard?

In the 1990s, a government agency released a database of medical visits, stripped of identifying information (names, addresses, social security numbers)

- But it did contain zip code, birth date, and gender.
- Researchers estimated that 87 percent of Americans are uniquely identifiable from this triplet.

Source: The Ethical Algorithm

Why Is Anonymization Hard?

Not sufficient to prevent unique identification of individuals

Name	Age	Gender	Zip Code	Smoker	Diagnosis
*	60–70	Male	191**	Y	Heart disease
*	60–70	Female	191**	N	Arthritis
*	60–70	Male	191**	Y	Lung cancer
*	60–70	Female	191**	N	Crohn's disease
*	60–70	Male	191**	Y	Lung cancer
*	***50–60***	***Female***	191**	N	HIV
*	50–60	Male	191**	Y	Lyme disease
*	50–60	Male	191**	Y	Seasonal allergies
*	***50–60***	***Female***	191**	N	Ulcerative colitis

Kearns & Roth, *The Ethical Algorithm*

From this (fictional) hospital database, if we know Rebecca is 55 years old and in this database, then we know she has 1 of 2 diseases.

Getting data about people is hard

We SOLVE tasks which are accessible:

- ✓ ImageNet
- ✓ MNIST
- ✓ CIFAR-10
- ✓ Librispeech
- ✓ WikiText-103

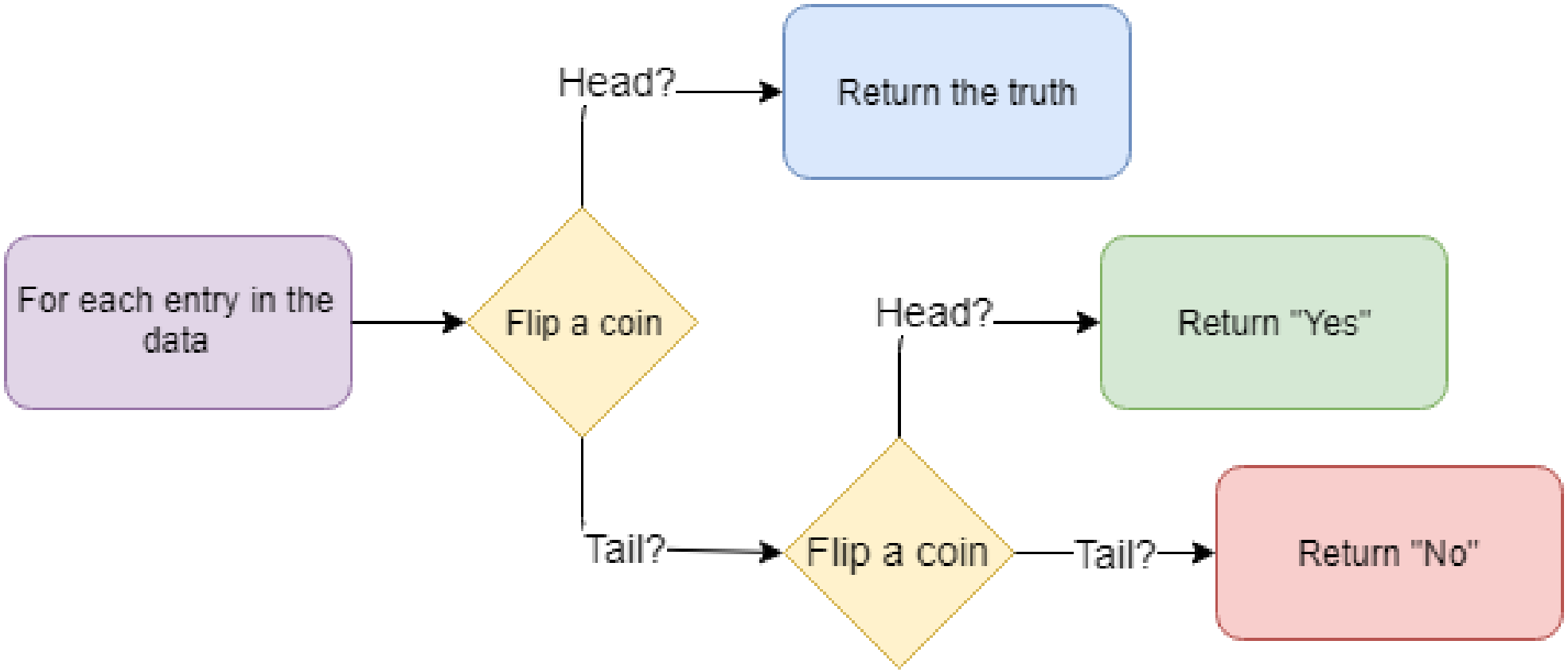
... but what about?

- Cancer
- Alzheimers
- Dementia
- Depression
- Anxiety
- ... the Common Cold?

Enterprise isolate their data

- Enterprises have a legal risk which prevents them from wanting to share their data set outside of their organization
- Enterprises have a competitive advantage to hang on to large datasets collected from or about their customers
- Some of the most important and personal issues in society cannot be addressed with ML because we do not have proper training data

Adding noise



CASE STUDY

When Private Data is Not Private

Large tech company Hooli spent the past year training an AI-powered health care program using personal information from one of the largest hospital systems BeHealthy in the U.S. Patients were not aware about it.

What happened:

- The tech company gave the BeHealthy hospital network access to a system for managing healthcare information called Project Morningale.
- In exchange, BeHealthy gave Hooli access to the medical records of up to 40 million patients.
- The effort triggered an investigation by U.S. privacy regulators.
- BeHealthy is the country's largest hospital system, that operates 250 hospitals.

More details

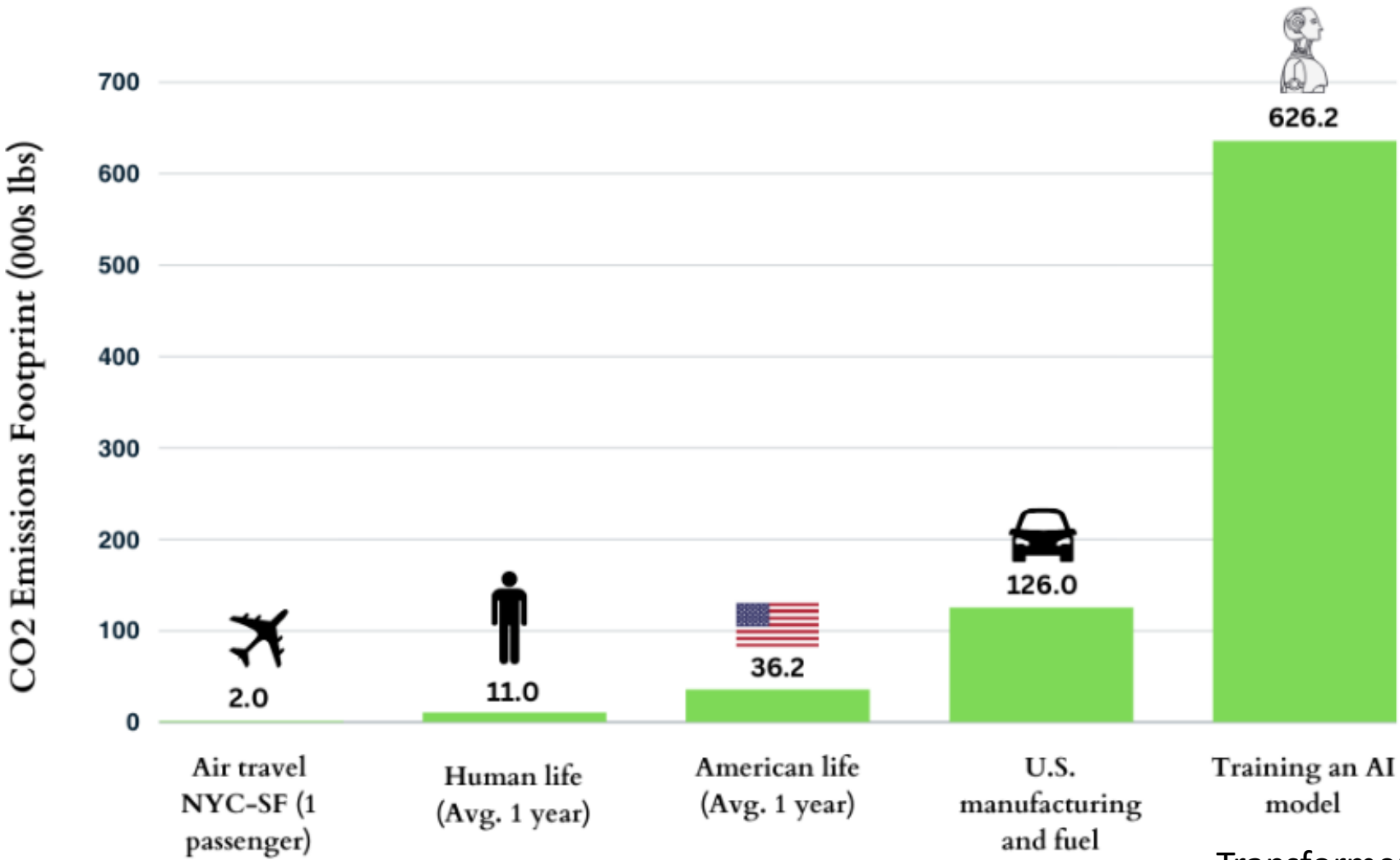
Hooli designed Project Morningale as an ML tool for matching patient information with healthcare decisions. Once trained, it would suggest treatment options or additional tests and highlight suggestions for special care based on a patient's history.

- The system would perform administrative tasks, such as reassigning doctors based on changes in the patient's condition or special needs. It would also enforce policies to prevent unlawful prescriptions and suggest ways to generate more income from patients.
- BeHealthy gave Hooli personal information (including patient names and addresses along with the names of patients' family members) as well as medical records such as lab results, diagnoses, prescriptions, and hospitalizations.
- BeHealthy didn't inform patients or doctors that it was sharing data with Hooli.
- At least 100 Hooli employees had access to the data.

From the point of view from a tech company Hooli, suggest steps that will ensure ethical approach for all involved parties, as well as help to build a robust AI system, that will help in healthcare.


Environmental Impact of AI

CO2 Emissions Benchmarks

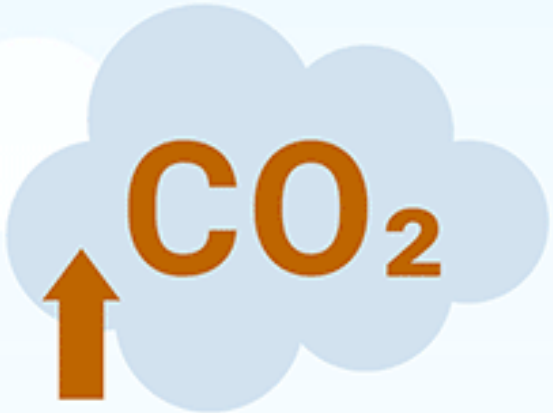


Transformer with 213M parameters

OWL ESG

Source: College of Information and Computer Sciences at University of Massachusetts Amherst

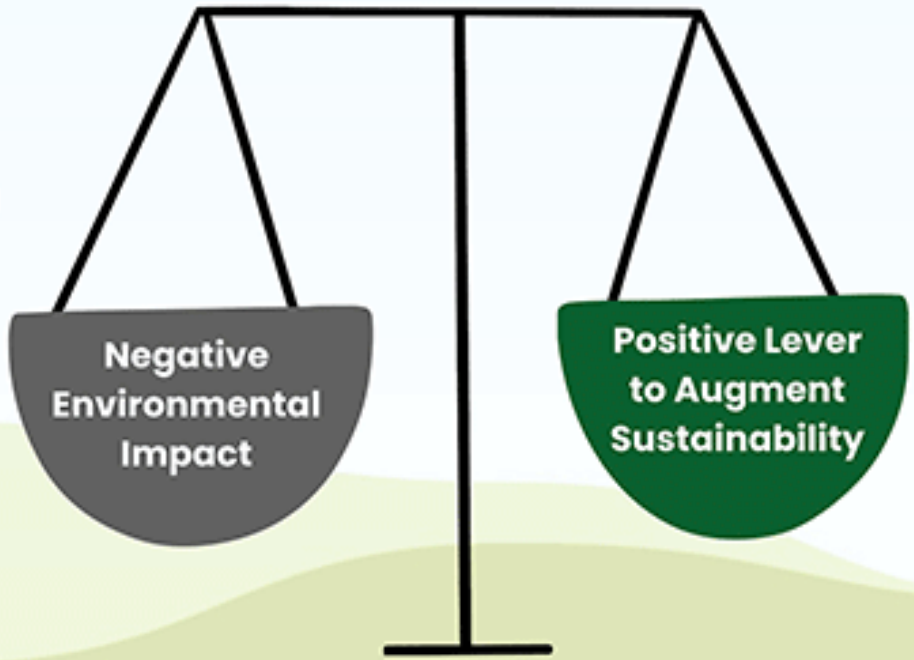
ENVIRONMENTAL IMPACT OF RADIOLOGY AI



AI Model Development and Deployment

Data Storage Requirements

Energy Source Choices



Scanner Operational Efficiency

Image Acquisition and Processing

Clinical Decision Support Tools

Opportunistic Screening and Analysis

Contrast Waste and Contamination

Patient and Radiology Workforce Scheduling