



University at Buffalo

Department of Computer Science
and Engineering

School of Engineering and Applied Sciences

ETHICS IN AI

Lecture 27

CSE 4/510: Reinforcement Learning

Instructor: Alina Vereshchaka

November 26, 2019

FACEBOOK'S ARTIFICIAL INTELLIGENCE ROBOTS SHUT DOWN AFTER THEY START TALKING TO EACH OTHER IN THEIR OWN LANGUAGE

'you i i i everything else'

Andrew Griffin | Monday 31 July 2017 17:10 | 89 comments



Click to follow
The Independent Tech

Facebook abandoned an experiment after two artificially intelligent programs appeared to be chatting to each other in a strange language only they understood.

The two chatbots came to create their own changes to English that made it easier for them to work – but which remained mysterious to the humans that supposedly look after them.

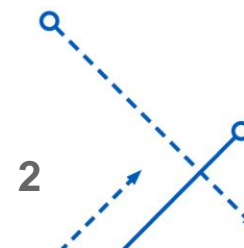
indy100 TRENDING



Report finds that Britain's most powerful 'more likely to be privately



Trump supporter says black people are attached to



FACEBOOK'S ARTIFICIAL INTELLIGENCE ROBOTS SHUT DOWN AFTER THEY START TALKING TO EACH OTHER IN THEIR OWN LANGUAGE

'you i i i everything else'

Andrew Griffin | Monday 31 July 2017 17:10 | 89 comments








[Click to follow The Independent Tech](#)

Facebook abandoned an experiment after two artificially intelligent programs appeared to be chatting to each other in a strange language only they understood.


The two chatbots came to create their own changes to English that made it easier for them to work – but which remained mysterious to the humans that supposedly look after them.



TRENDING

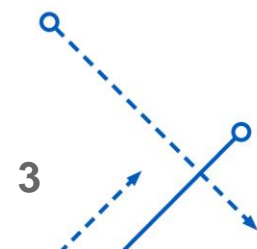


Report finds that Britain's most powerful 'more likely to be privately



Trump supporter says black neonele are attached to

But to whom they were fearing?
What could happen if they let go the program as it is?



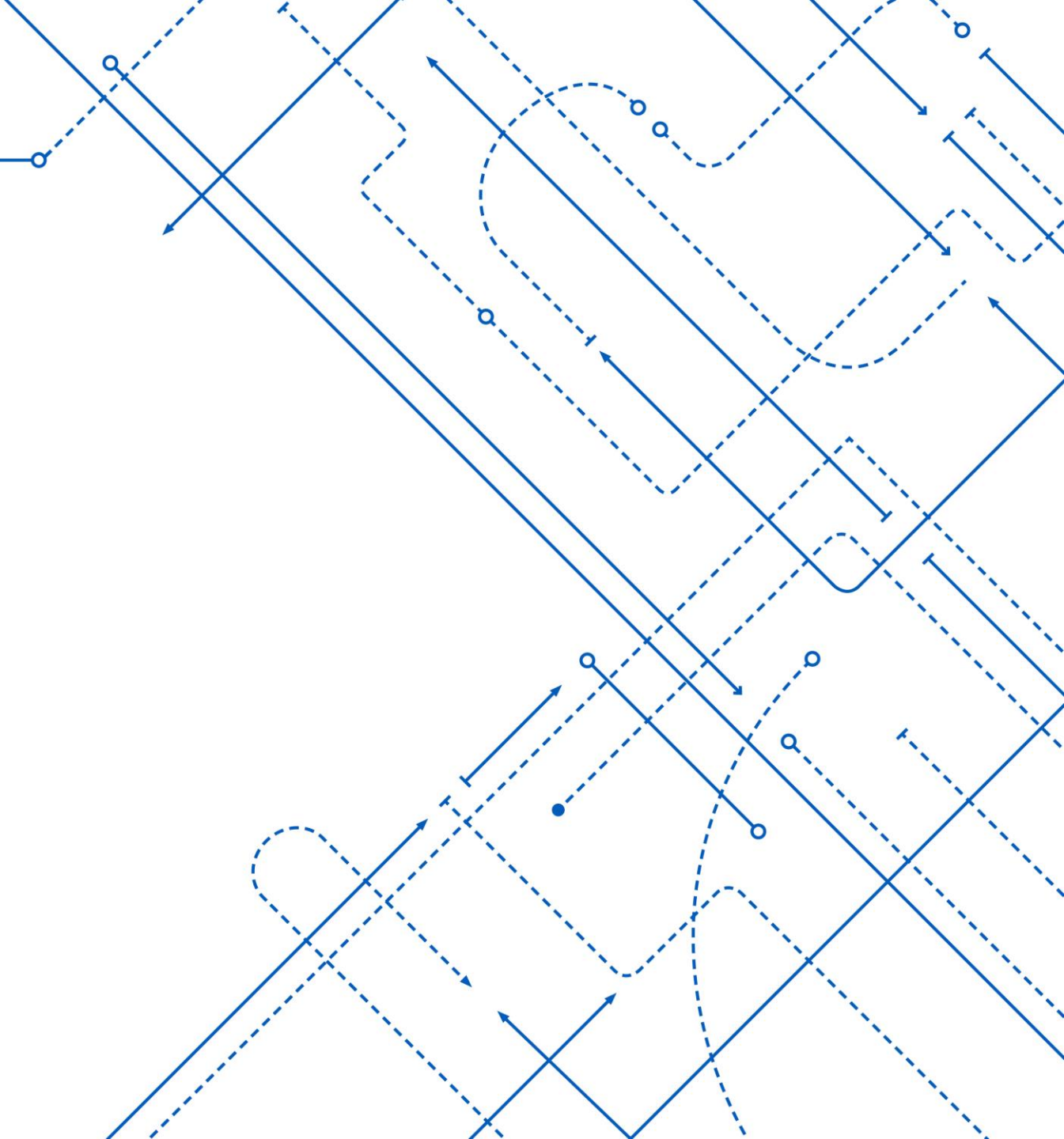


University at Buffalo

Department of Computer Science
and Engineering

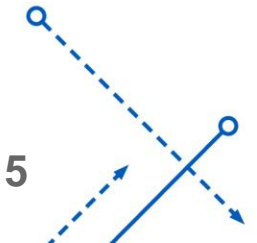
School of Engineering and Applied Sciences

GOAL OF AI



Question #1

What is the goal of AI?



What is the Goal of AI?



OECD's* Principles on AI:

“AI should benefit people and the planet”

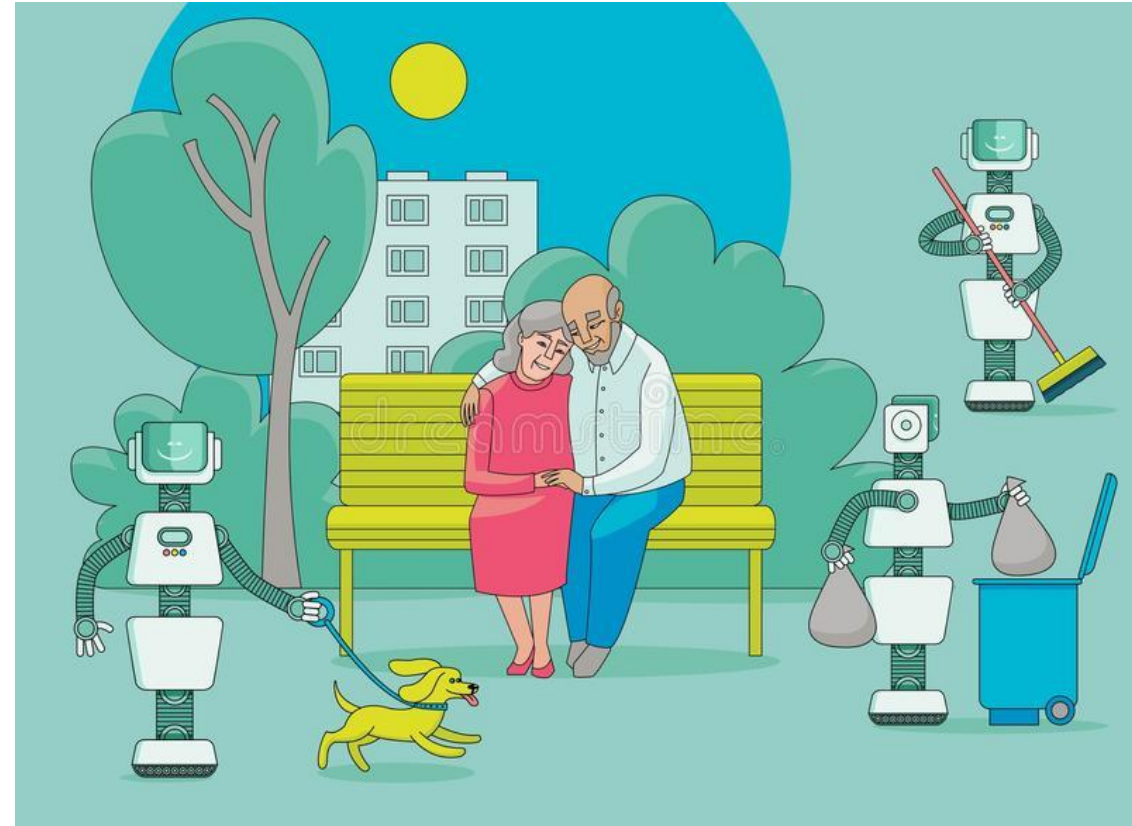
***Organization for Economic Co-operation and Development (OECD)** is an intergovernmental economic organization with 36 member countries to stimulate economic progress and world trade.

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

OECD's Principles on AI

INCLUSIVE GROWTH, SUSTAINABLE DEVELOPMENT AND WELL-BEING

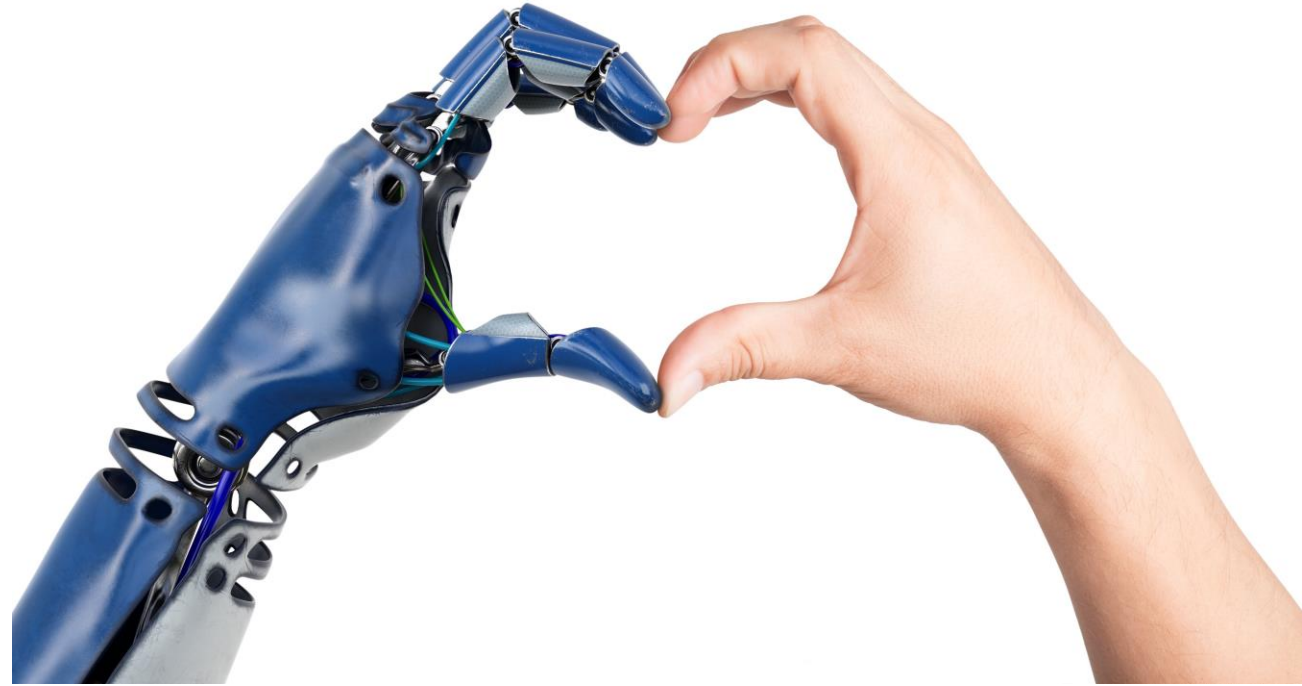
Trustworthy AI in pursuit of beneficial outcomes for people and the planet.



OECD's Principles on AI

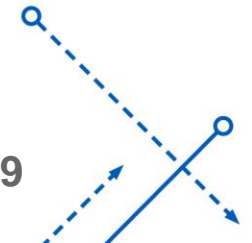
HUMAN-CENTRED VALUES AND FAIRNESS

AI should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle.



Ethics in AI

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

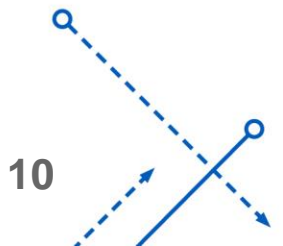


Ethics in AI

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



Ethics in AI

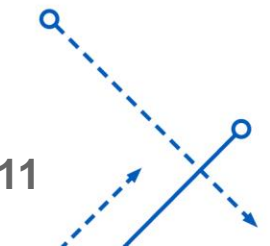
Humans are intelligent to the extent that **our** actions can be expected to achieve **our** objectives.

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

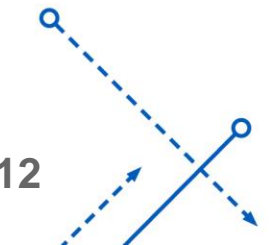
- Control theory: minimize cost function
- Economics: maximize expected utility
- Operations research: maximize sum of rewards
- Statistics: minimize loss function
- AI: all of the above, plus logically defined goals

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

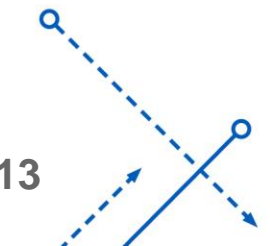
We need machines to be provably beneficial



1. The machine's **only objective** is to **maximize the realization of human preferences**

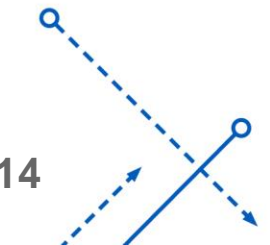


1. The machine's **only objective** is to **maximize the realization of human preferences**
2. The robot is initially **uncertain** about what those preferences are



Ethics in AI

1. The machine's **only objective** is to **maximize the realization of human preferences**
2. The robot is initially **uncertain** about what those preferences are
3. **Human behavior** provides evidence about **human preferences**



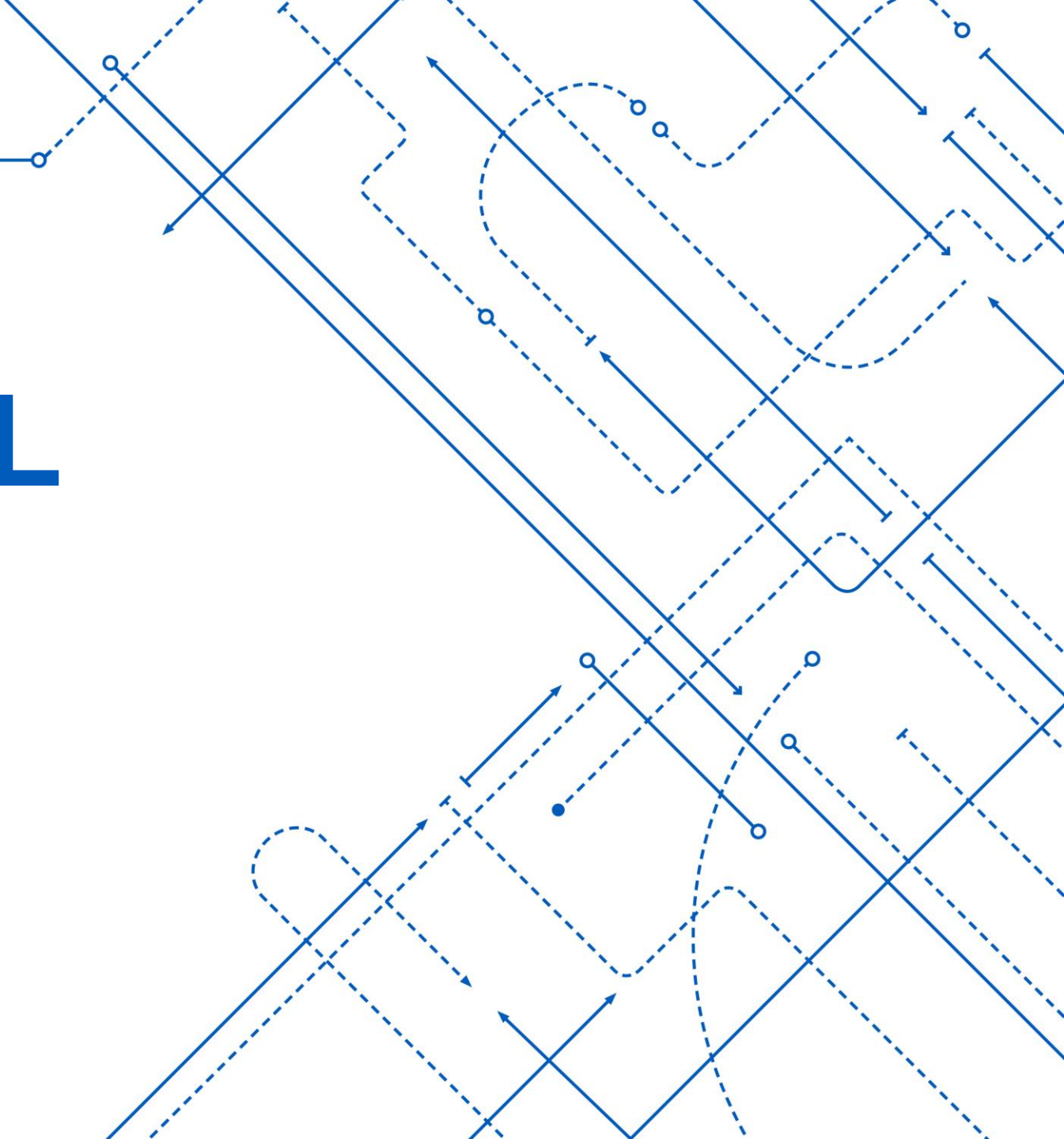


University at Buffalo

Department of Computer Science
and Engineering

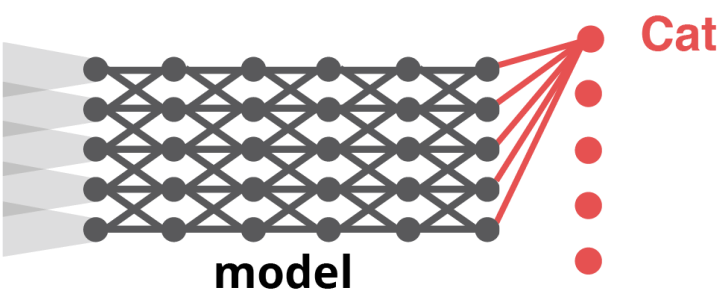
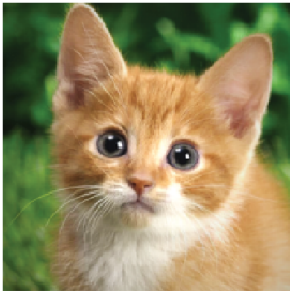
School of Engineering and Applied Sciences

ADVERSARIAL ATTACKS

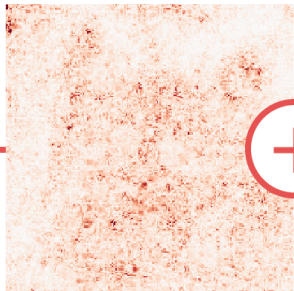
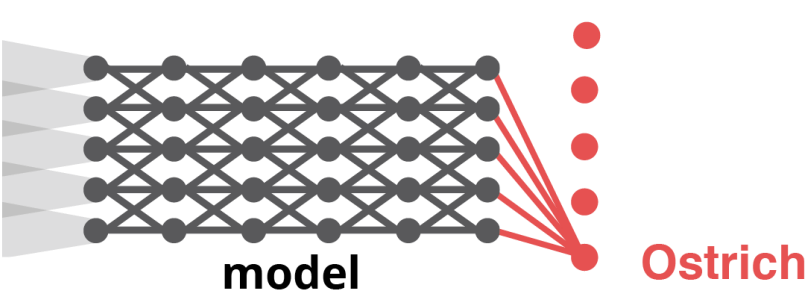
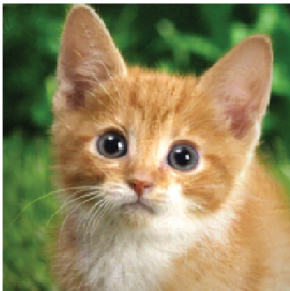


Adversarial Attacks

Original image



Adversarial image



(small) adversarial perturbation
created by **attack**



Adversarial Attacks



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

Adversarial Attacks



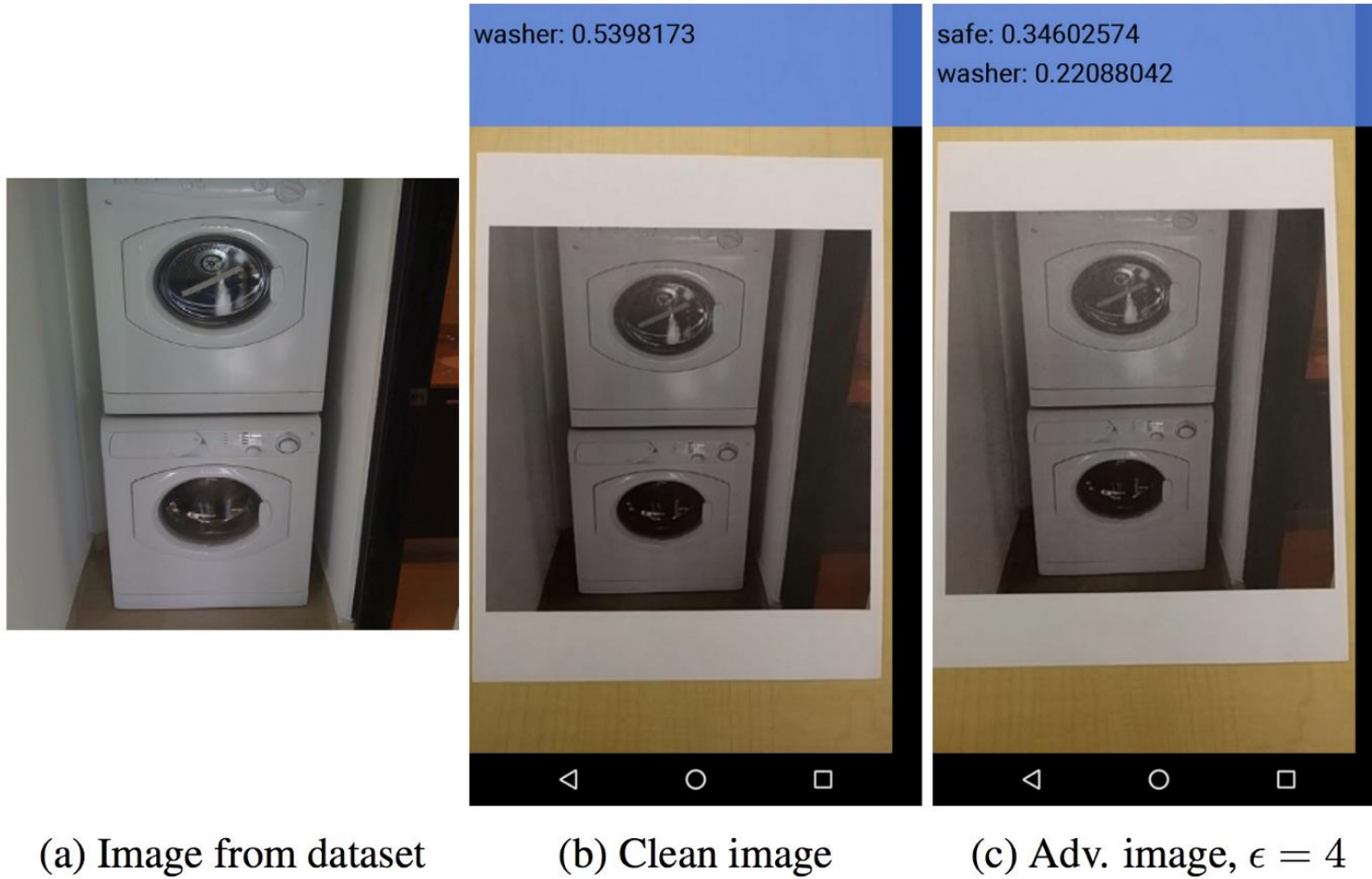
(a) Image from dataset



(b) Clean image

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

Adversarial Attacks



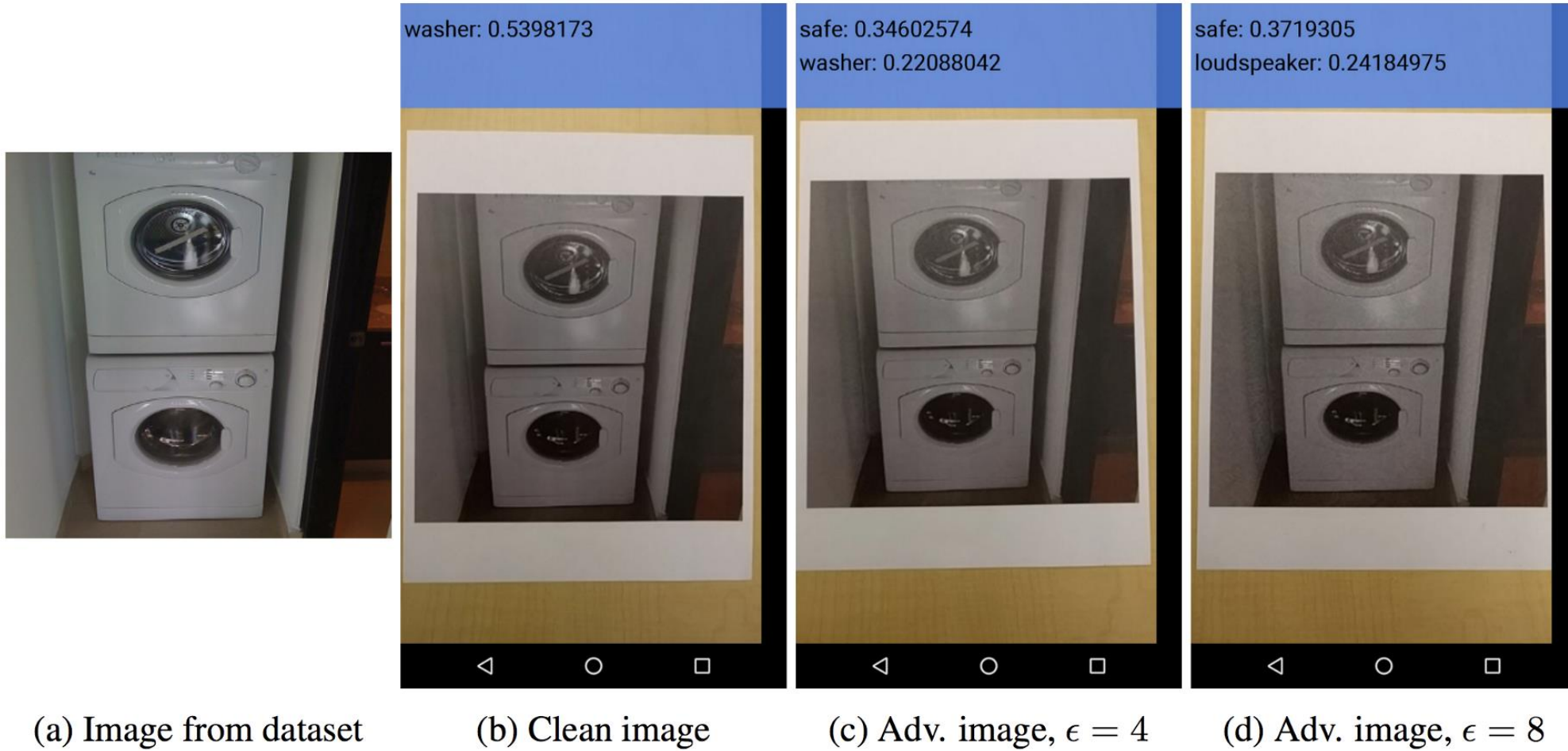
(a) Image from dataset

(b) Clean image

(c) Adv. image, $\epsilon = 4$

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

Adversarial Attacks

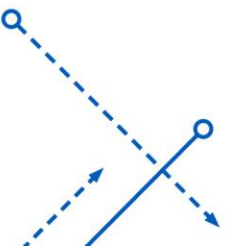


Source: [Adversarial Examples in the Physical World](#). Kurakin et al, ICLR 2017.

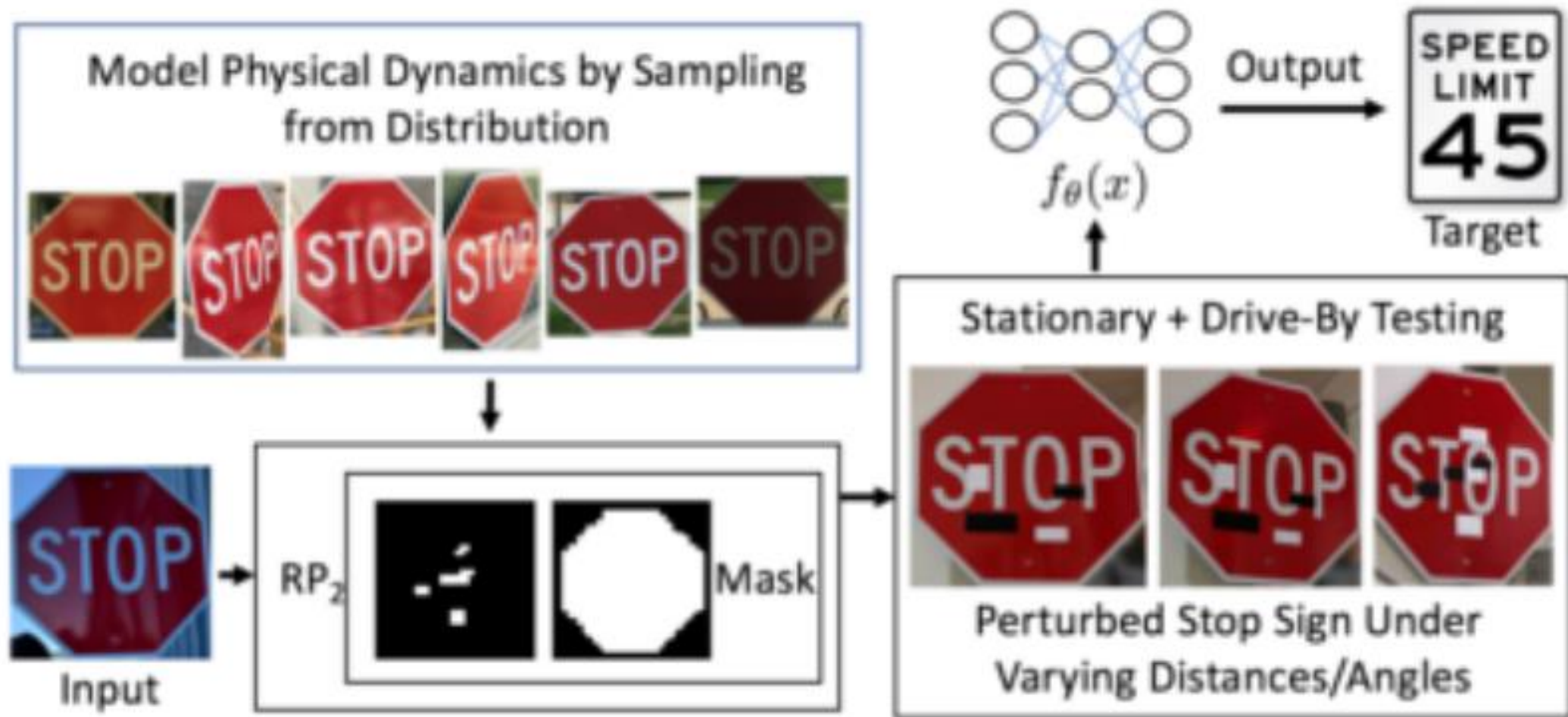
Adversarial Attacks



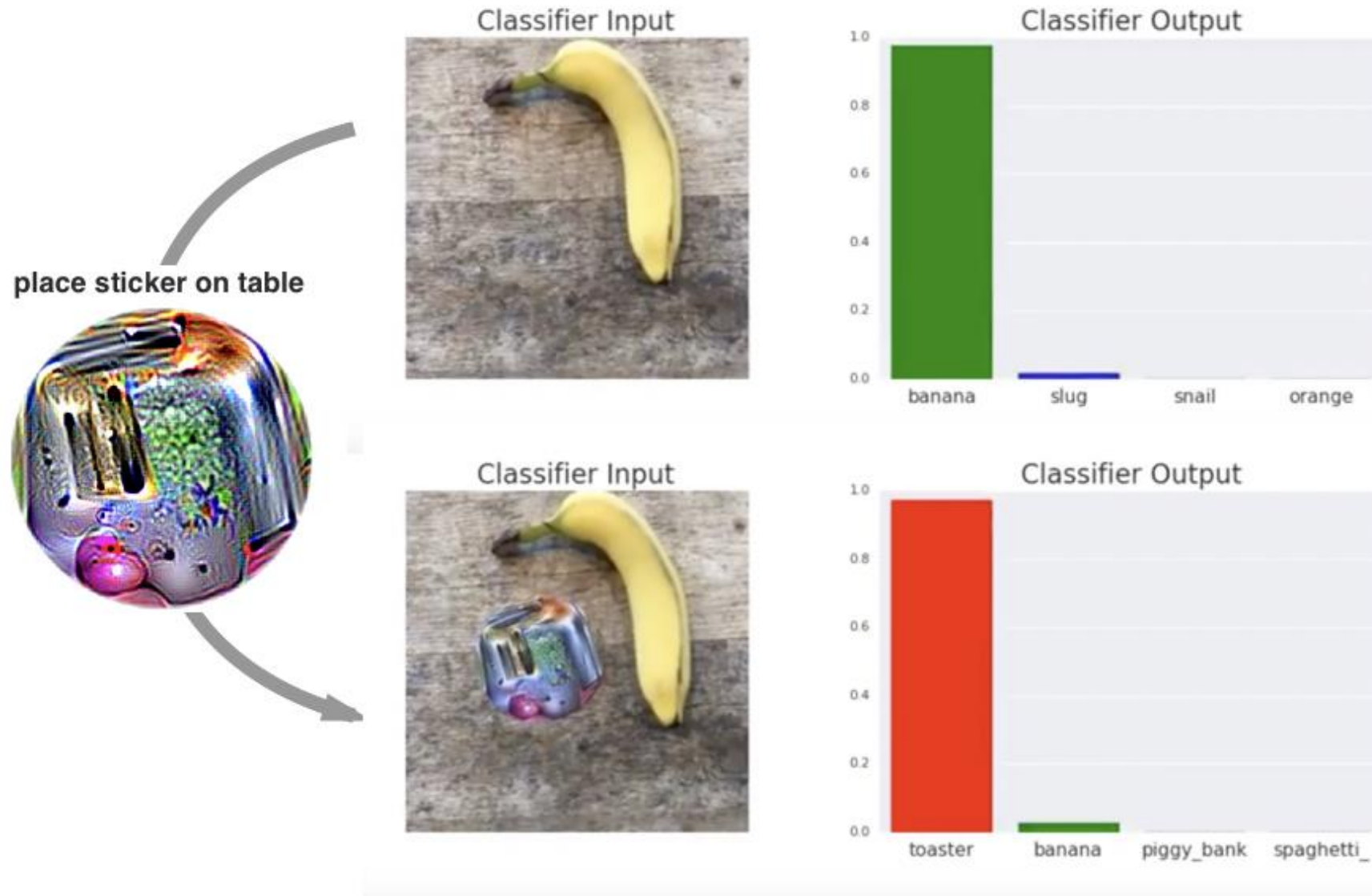
Figure 1: The left image shows real graffiti on a Stop sign, something that most humans would not think is suspicious. The right image shows our a physical perturbation applied to a Stop sign. We design our perturbations to mimic graffiti, and thus “hide in the human psyche.”



Adversarial Attacks

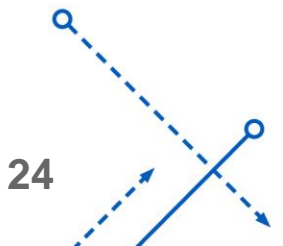


Adversarial Attacks



Question #2: Adversarial Attacks

How can we prevent adversarial attacks?



How can we prevent adversarial attacks?

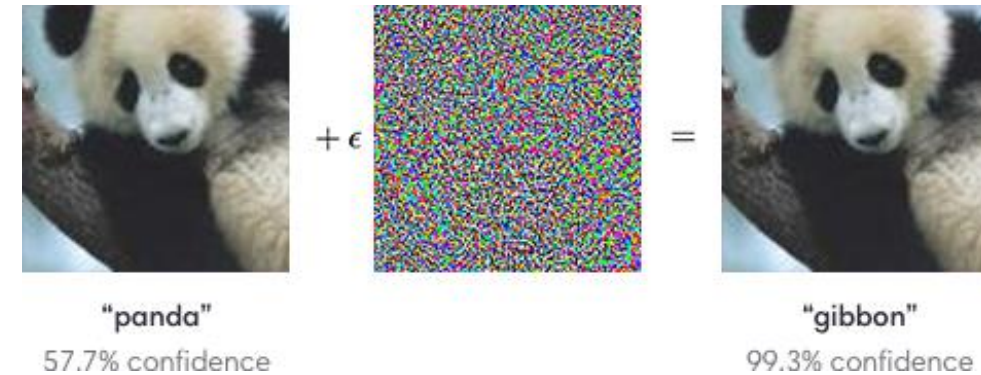
Possible solutions

- **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.



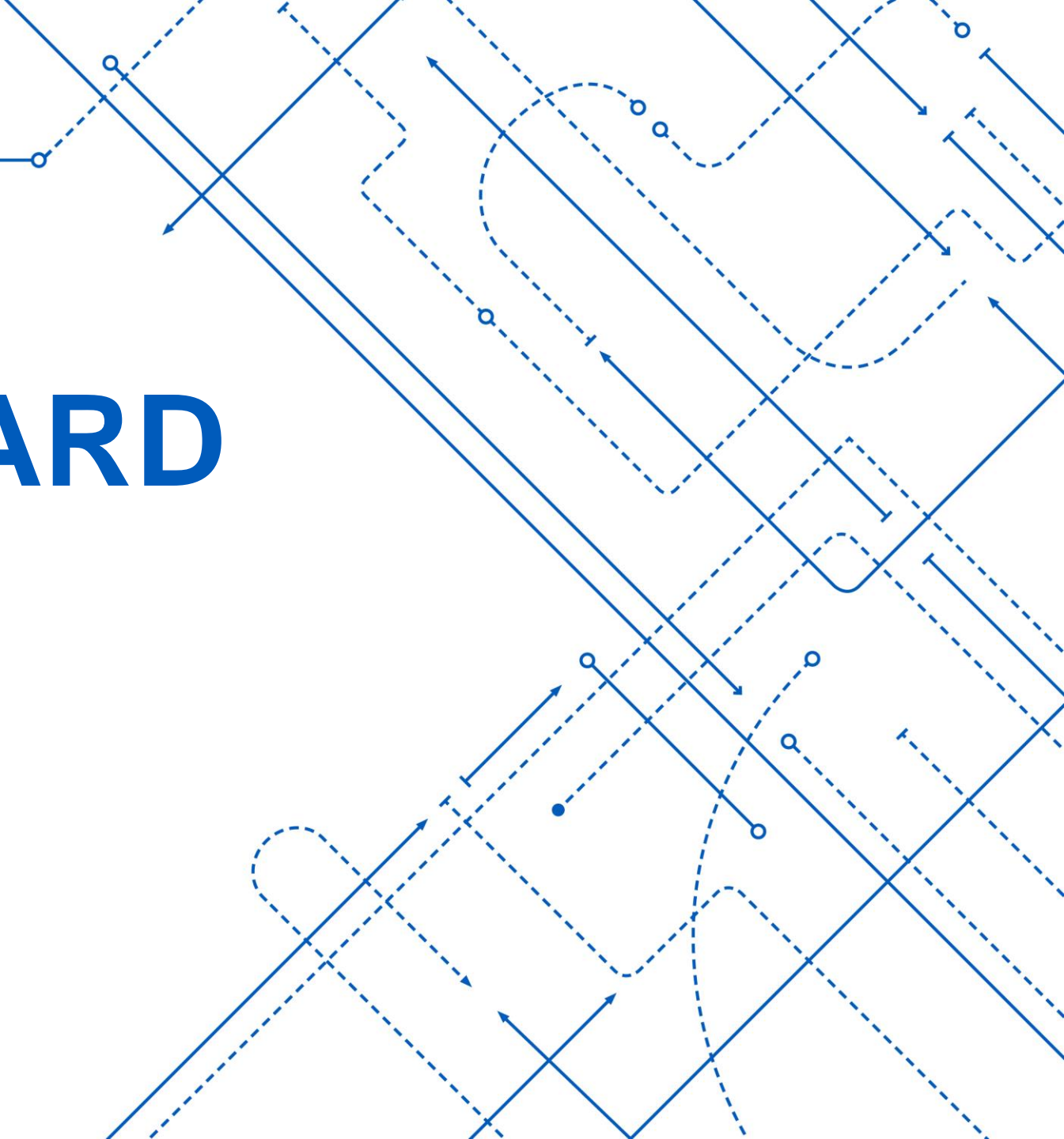


University at Buffalo

Department of Computer Science
and Engineering

School of Engineering and Applied Sciences

FAULTY REWARD FUNCTIONS



Faulty Reward Functions



Source: <https://openai.com/blog/faulty-reward-functions/>

Question #4: Faulty Reward Functions

How can we avoid faulty rewards?



Faulty Reward Functions

Possible Solutions

- Learning from demonstrations
- Incorporate human feedback by evaluating the quality of episodes
- Use transfer learning to train on many similar games, and infer a “common sense” reward function for this game



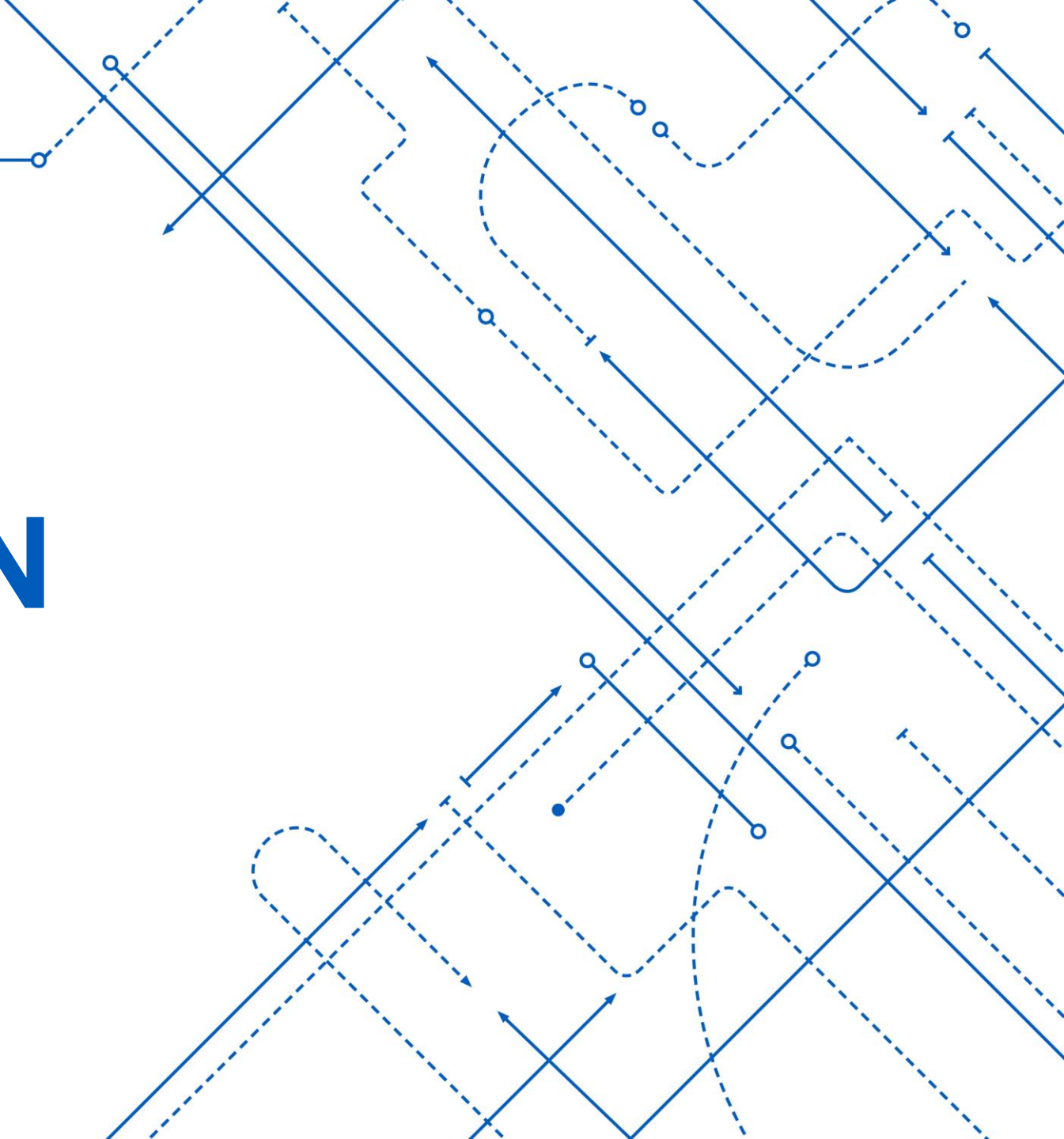


University at Buffalo

Department of Computer Science
and Engineering

School of Engineering and Applied Sciences

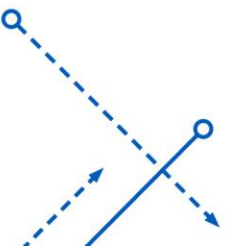
UNSAFE EXPLORATION



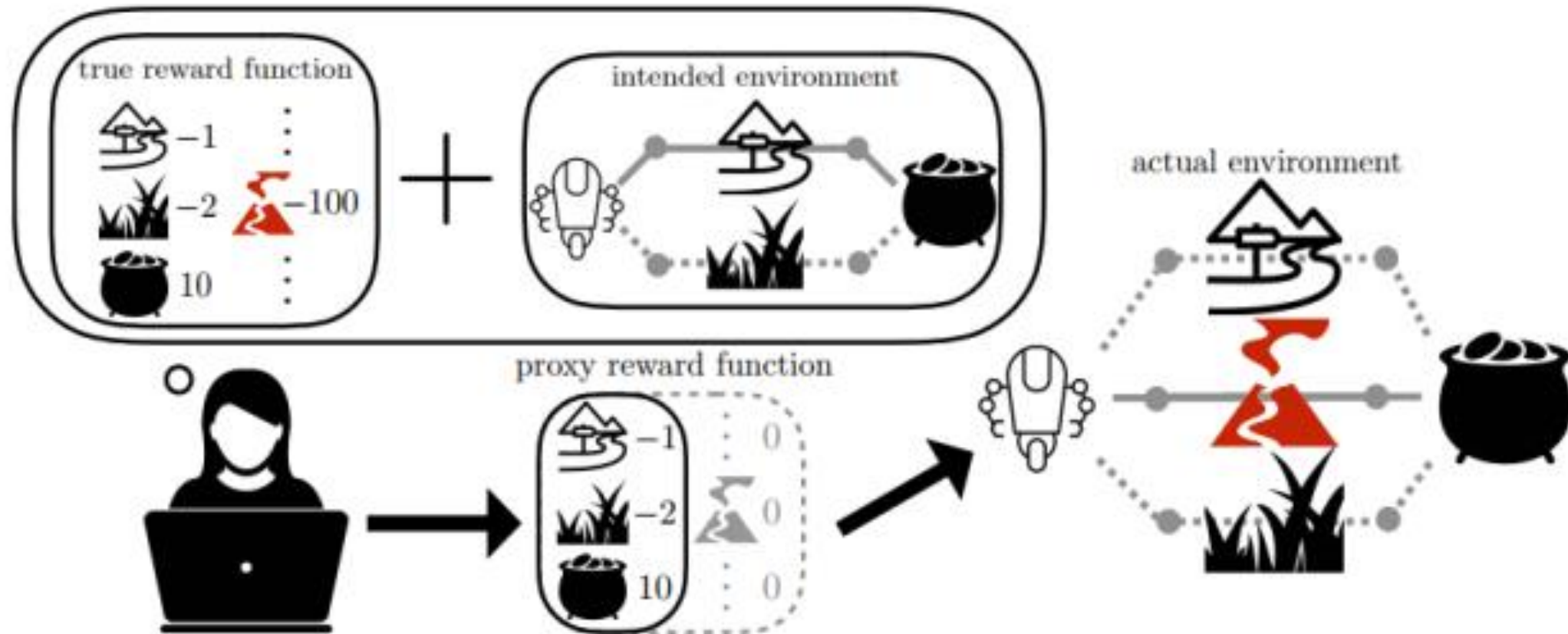
Safe Reinforcement Learning

Safe Reinforcement Learning can be defined as the process of learning policies that maximize the expectation of the return in problems in which it is important to **ensure reasonable system performance** and/or **respect safety constraints** during the learning and/or deployment processes.

- Garcia et al 2015



Unsafe exploration

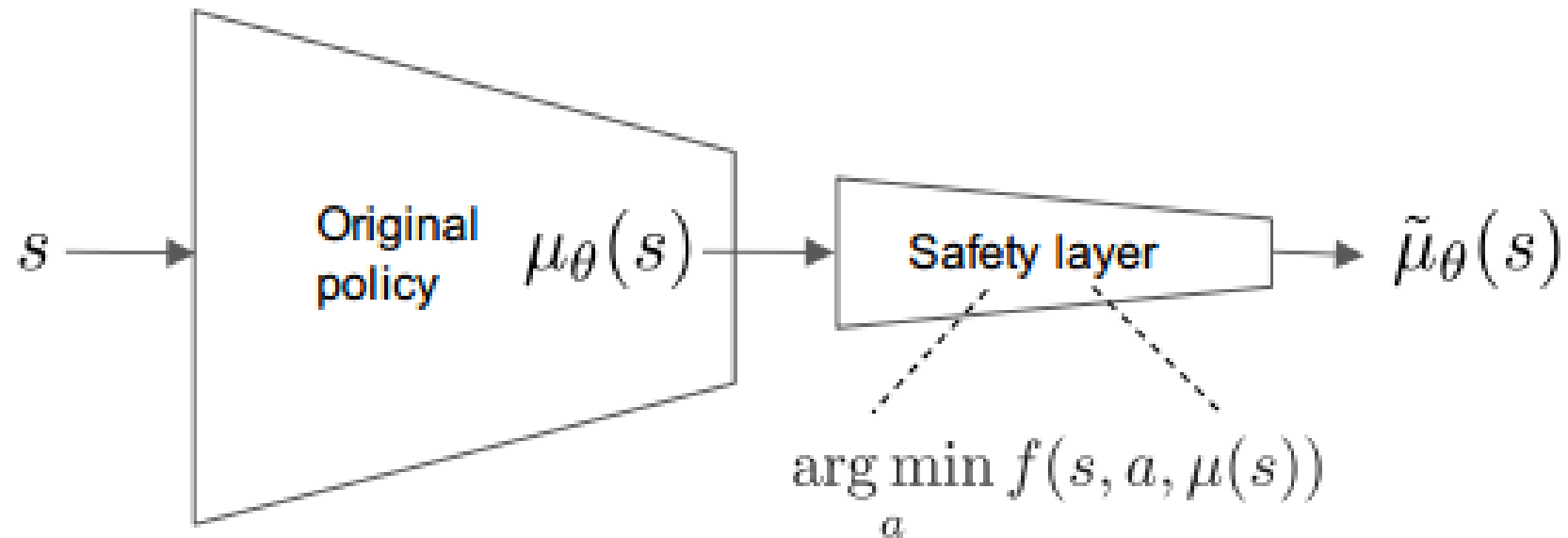


Safe exploration

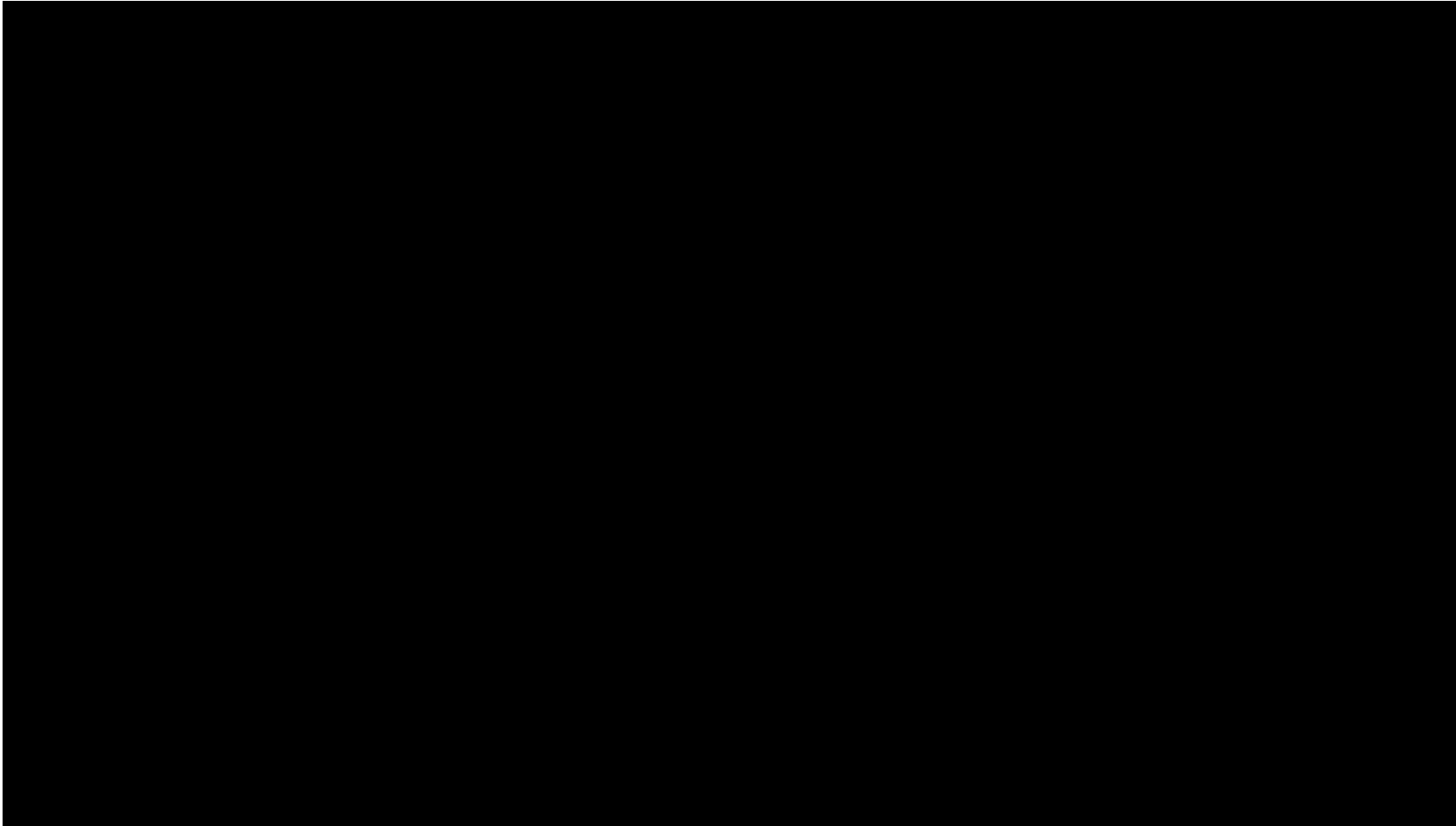


- The policy gradient (dark blue arrow) points in the direction that increases reward
- The optimal step before considering safety constraints (dotted blue arrow) lies on the edge of the KL trust region (blue oval)
- But to step in the constraint-satisfying area (light green half-space). The step is adjusted (to the dark green arrow), so that reward is increased as much as possible but also staying safe.

Safe exploration



Unsafe exploration



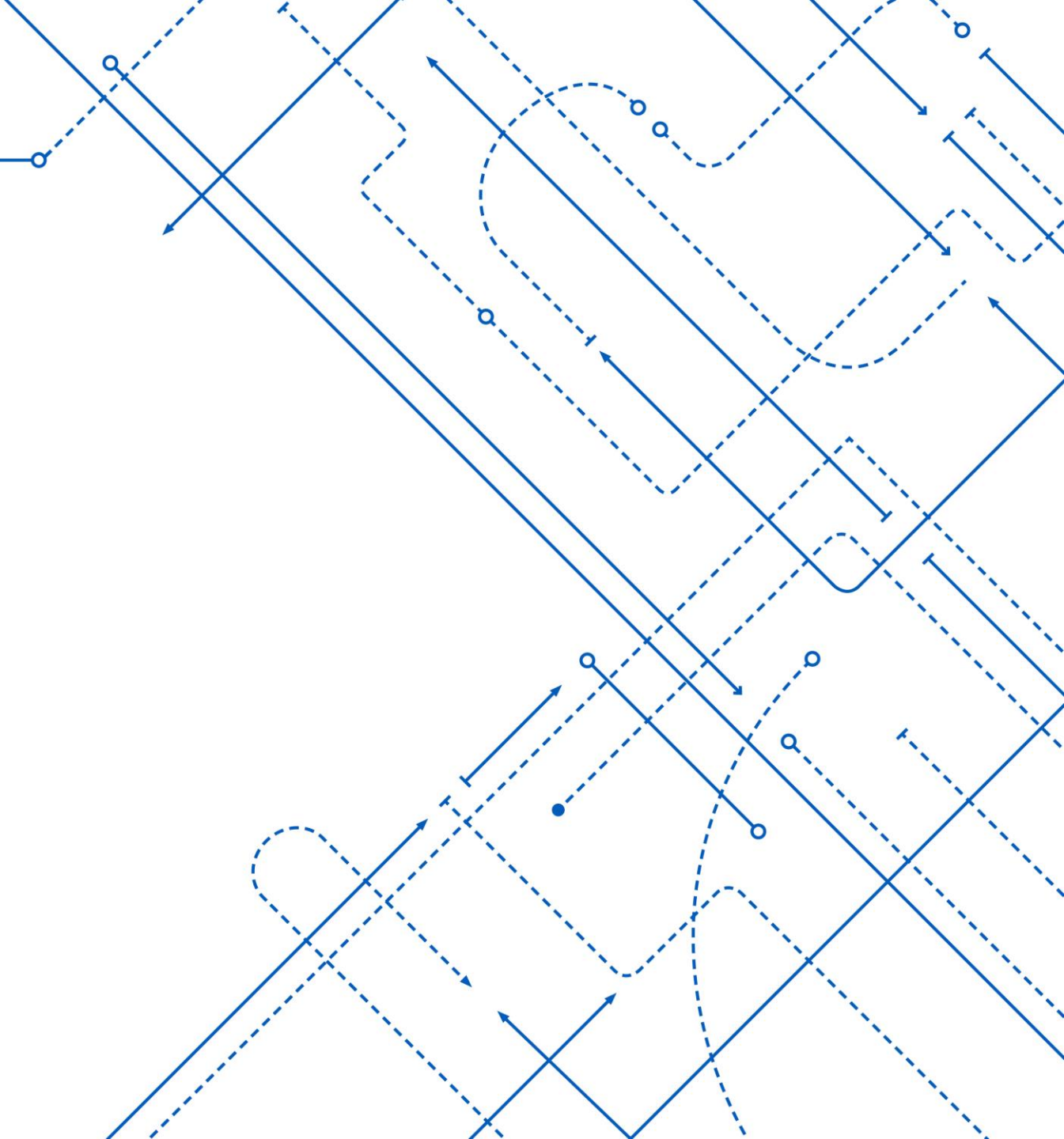


University at Buffalo

Department of Computer Science
and Engineering

School of Engineering and Applied Sciences

AI MORALITY



Morality in human autonomy is a complex philosophical problem. Do the right thing



AI Morality

Morality in human autonomy is a complex philosophical problem. **Do the right thing**

Morality in machine autonomy is, for the time being, an engineering problem. **Do what you are told.**

Question #5: AI Morality

How to provide the behavioral constraints to the agent?

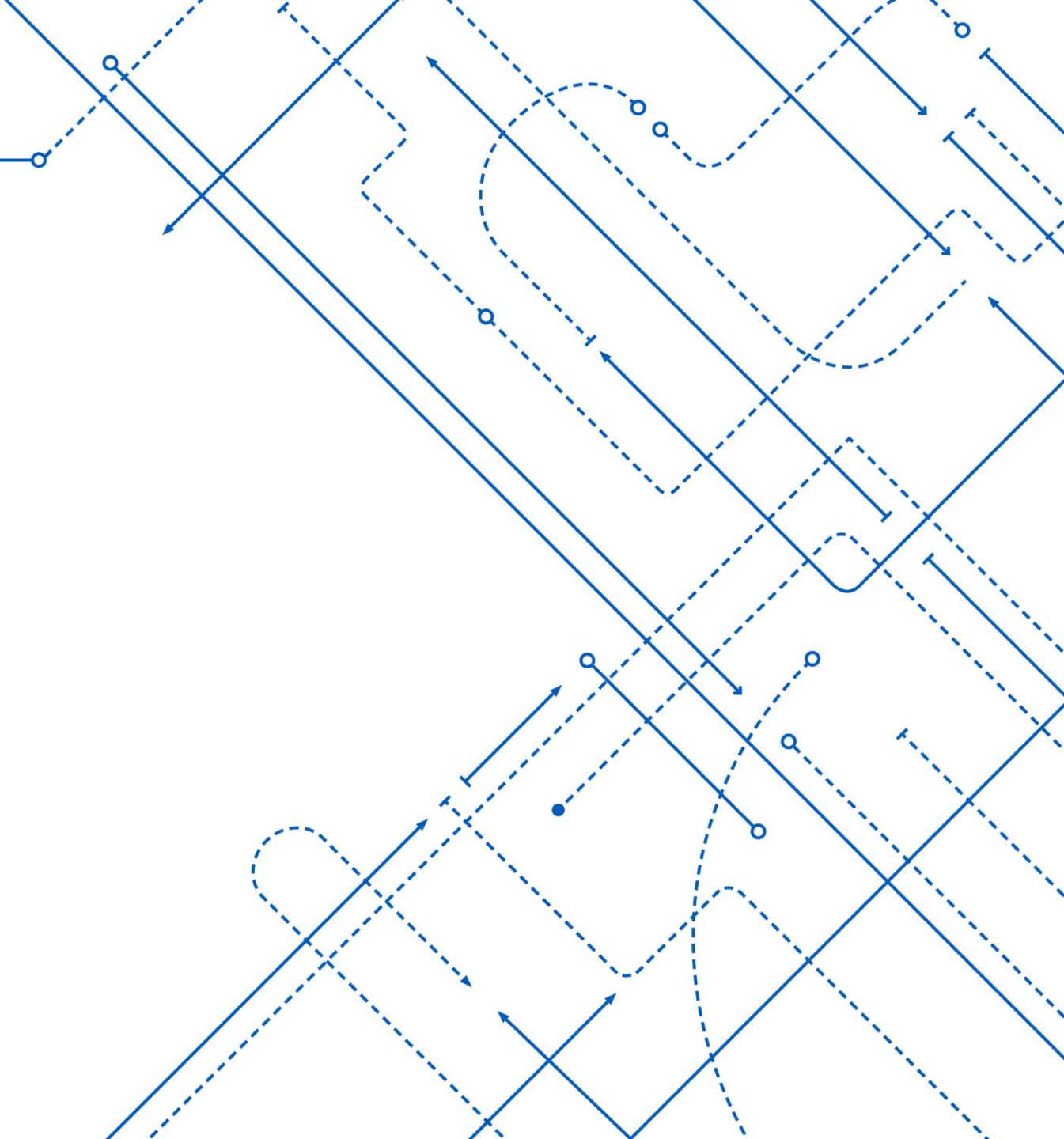


University at Buffalo

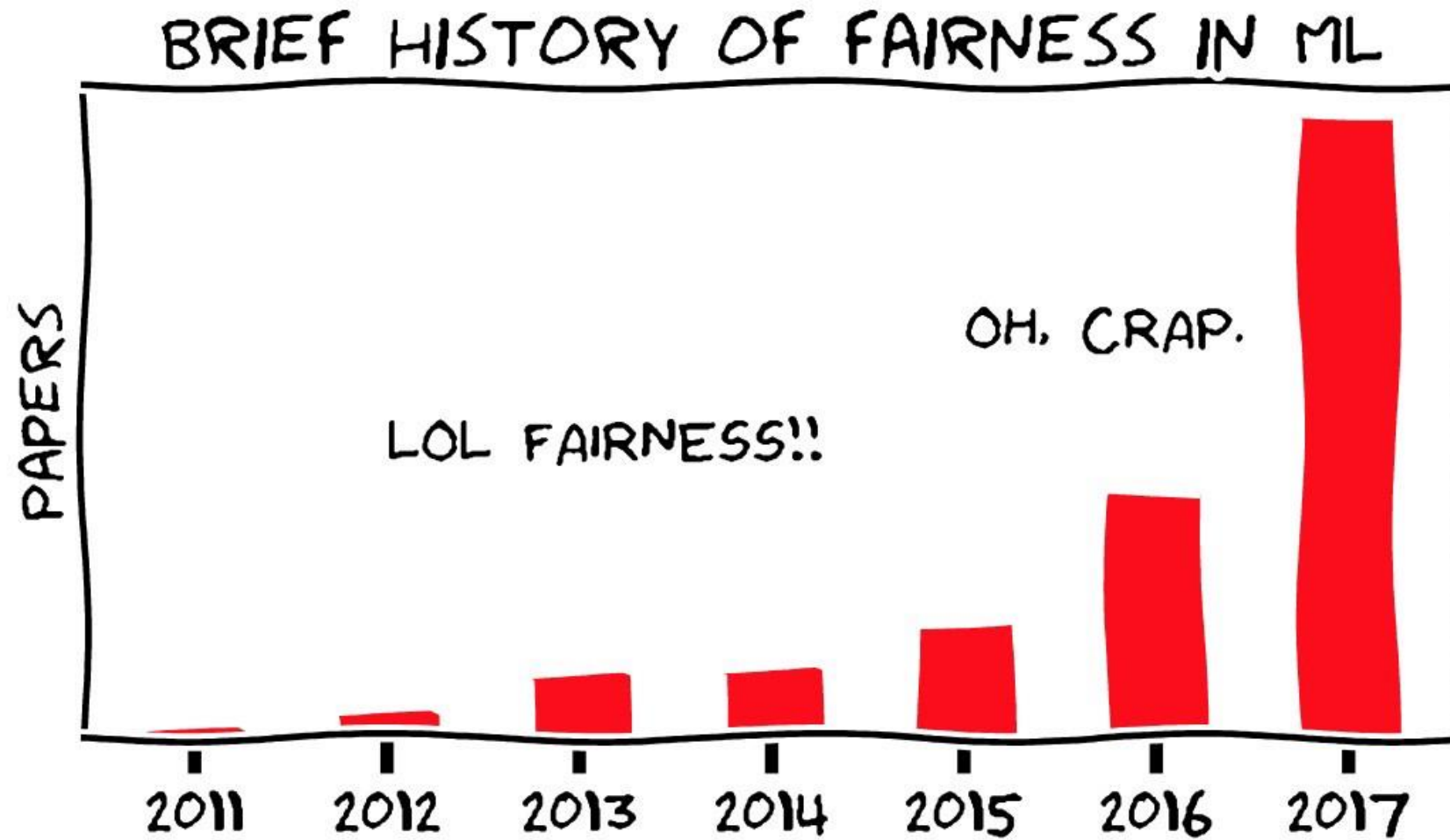
Department of Computer Science
and Engineering

School of Engineering and Applied Sciences

FAIRNESS



Fairness in AI



What is Bias/Fairness?

“**Bias**. When scientific or **technological decisions** are based on a narrow set of systemic, structural or **social concepts and norms**, the resulting technology can **privilege certain groups** and harm others.” – Nature comment

What is Bias/Fairness?

We live in a biased society, so it's inevitable that data collected about that society will be biased: inherent bias, test data, feedback, proxies.



| COOKING | |
|---------|---------|
| ROLE | VALUE |
| AGENT | WOMAN |
| FOOD | PASTA |
| HEAT | STOVE |
| TOOL | SPATULA |
| PLACE | KITCHEN |



| COOKING | |
|---------|---------|
| ROLE | VALUE |
| AGENT | WOMAN |
| FOOD | FRUIT |
| HEAT | ∅ |
| TOOL | KNIFE |
| PLACE | KITCHEN |



| COOKING | |
|---------|---------|
| ROLE | VALUE |
| AGENT | WOMAN |
| FOOD | MEAT |
| HEAT | STOVE |
| TOOL | SPATULA |
| PLACE | OUTSIDE |



| COOKING | |
|---------|---------|
| ROLE | VALUE |
| AGENT | WOMAN |
| FOOD | ∅ |
| HEAT | STOVE |
| TOOL | SPATULA |
| PLACE | KITCHEN |



| COOKING | |
|---------|---------|
| ROLE | VALUE |
| AGENT | MAN |
| FOOD | ∅ |
| HEAT | STOVE |
| TOOL | SPATULA |
| PLACE | KITCHEN |



Prediction Fails Differently for Black Defendants

| | WHITE | AFRICAN AMERICAN |
|---|-------|------------------|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 47.7% | 28.0% |

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Source:

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Question #6: Biases in Algorithms

How can we make sure algorithms are fair, especially when they are privately owned by corporations, and not accessible to public scrutiny?

How can we balance openness and intellectual property?





University at Buffalo

Department of Computer Science
and Engineering

School of Engineering and Applied Sciences

TRANSPARENCY OF ALGORITHMS

Question #7: Transparency of Algorithms

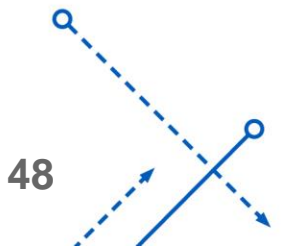
How can we balance the need for more accurate algorithms with the need for transparency towards people who are being affected by these algorithms?

If necessary, are we willing to sacrifice accuracy for transparency, as Europe's new General Data Protection Regulation may do?

Question #8: Supremacy of Algorithms

If we start trusting algorithms to make decisions, who will have the final word on important decisions?

Will it be humans, or algorithms?



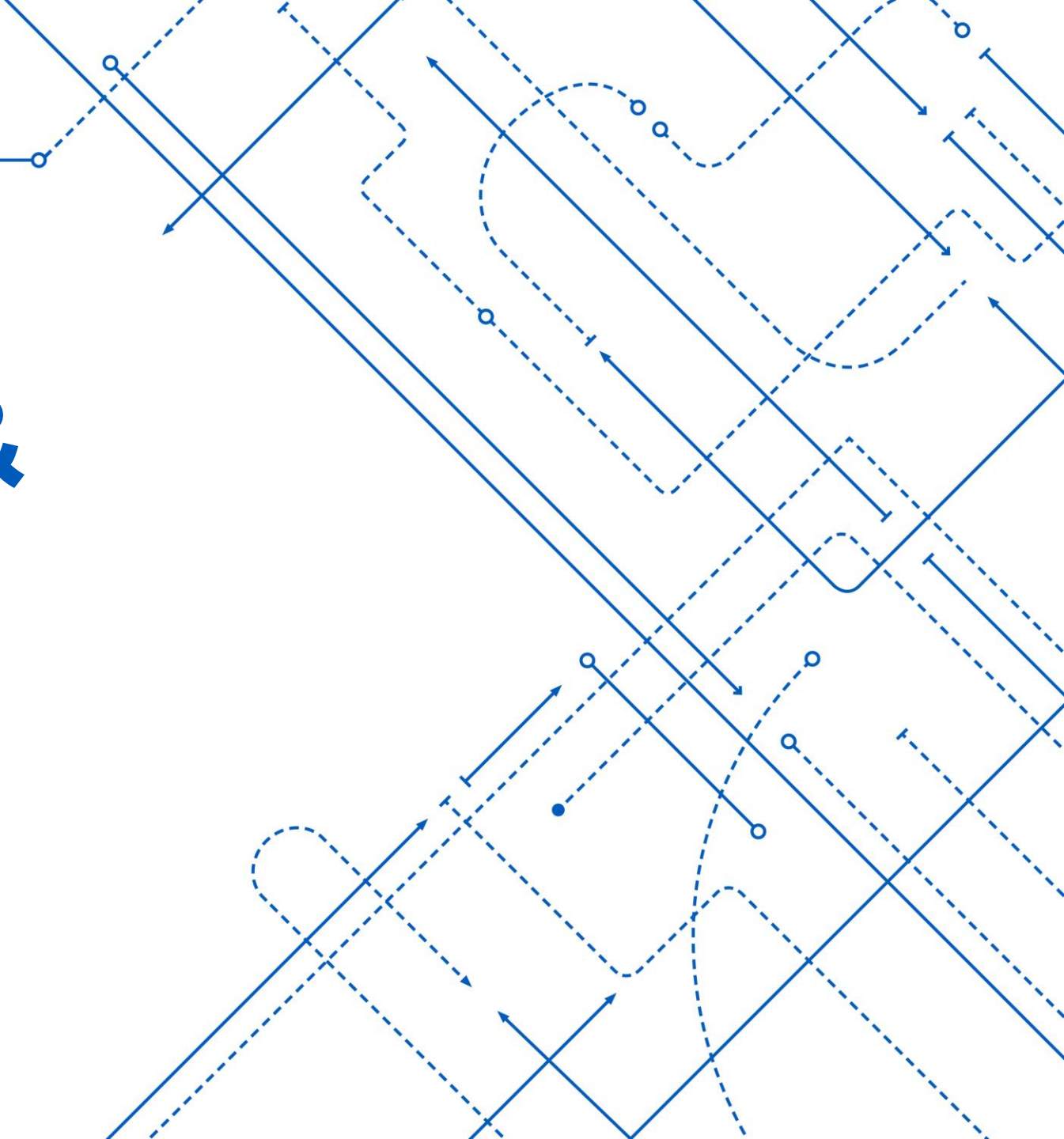


University at Buffalo

Department of Computer Science
and Engineering

School of Engineering and Applied Sciences

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

AI Text Generator – GPT-2 by OpenAI

Q: What should I do today?

A: Go. Read a book, look out for the train, check out the sunset, enjoy all the food in your town. If you're lucky, you might even notice a beautiful lake or waterfall! When it's your turn for the day, take a leisurely stroll to the water's edge, look at the beautiful mountains and lakes, and get to know the locals.

Scientist have discovered a new type of moose, dubbed the Ultra Moose. It stands 50 feet tall, has six pairs of muscular arms, and weighs between 200 and 250 pounds. The Ultra Moose is so large, Dr. Jekyll and Mr. Hyde might not be able to get close enough to get their teeth in its neck. The scientist named this new species Ultra Mouse for its unique abilities.

AI Text Generator – GPT-2 by OpenAI

SYSTEM PROMPT
(HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL
COMPLETION
(MACHINE-
WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Question #9: AI Synthetic World

How can we differentiate between real/generated data?



Question #10: Fake News & Fake Videos

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

How can we slow the spread of false information, and who will get to decide which news count as “true”?



University at Buffalo

Department of Computer Science
and Engineering

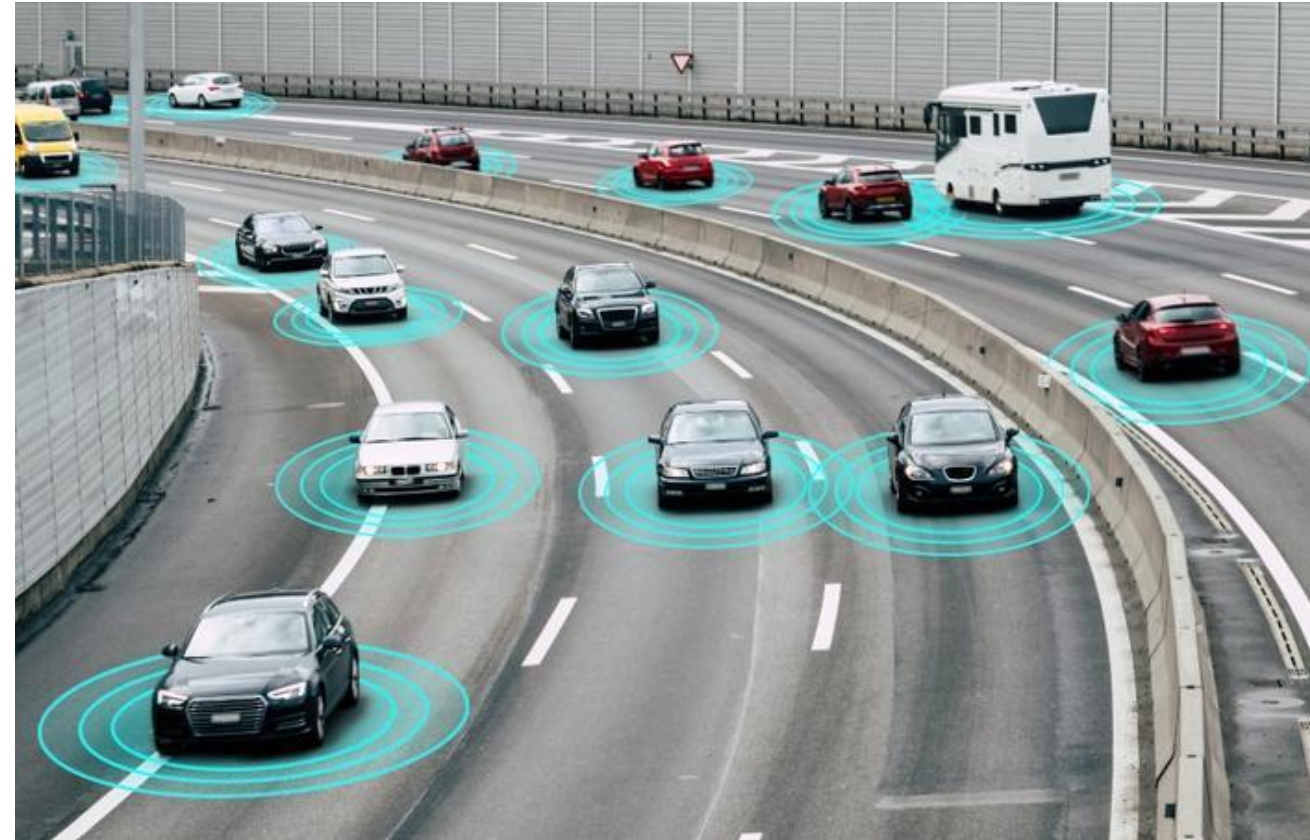
School of Engineering and Applied Sciences

SELF-DRIVING CARS



Self-driving Cars

- **10 million** autonomous vehicles will hit the roads by 2020
- **In 10 years** fully autonomous vehicles will be the norm
- AVs will generate a **\$7 trillion** annual revenue stream by 2050
- Widespread adoption of AVs could lead to a **90% reduction in vehicle crashes**



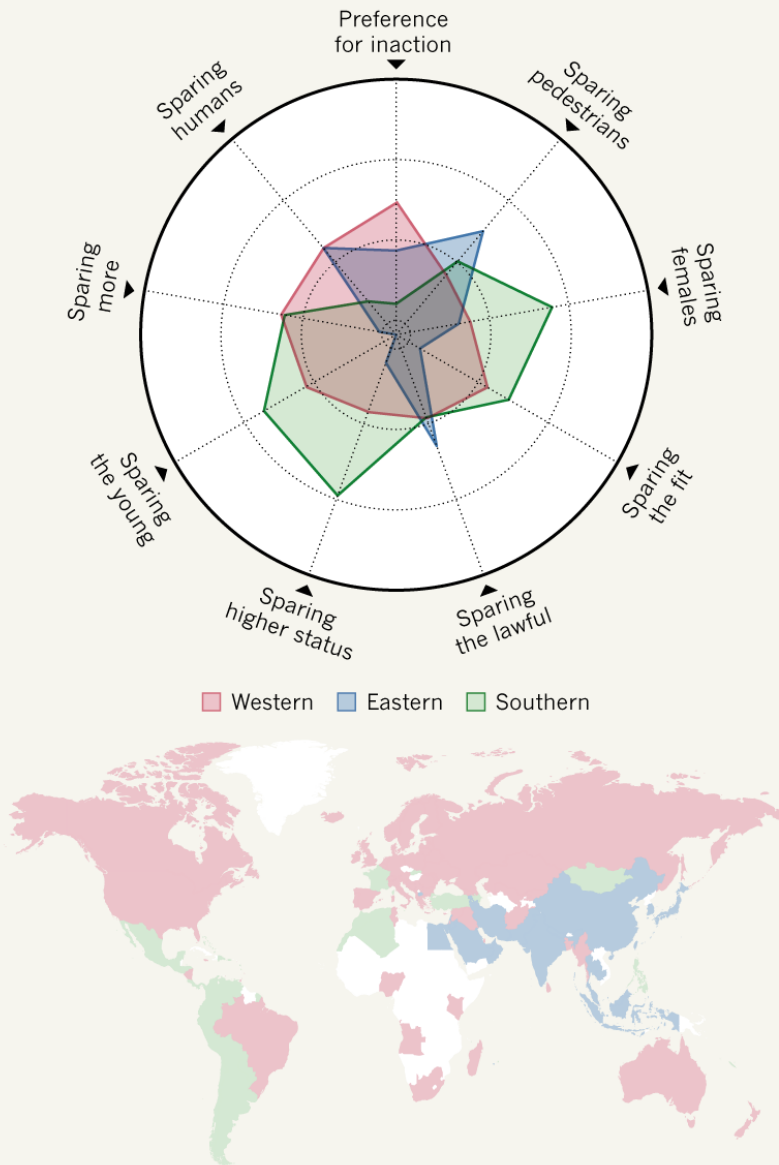
Self-driving Cars

When a driver slams on the brakes to avoid hitting a pedestrian crossing the road illegally, she is making a moral decision that shifts risk from the pedestrian to the people in the car.



MORAL COMPASS

A survey of 2.3 million people worldwide reveals variations in the moral principles that guide drivers' decisions. Respondents were presented with 13 scenarios, in which a collision that killed some combination of passengers and pedestrians was unavoidable, and asked to decide who they would spare. Scientists used these data to group countries and territories into three groups based on their moral attitudes.



- Authors analyzed answers from people in the 130 countries with at least 100 respondents
- Western group showed a stronger preference for sacrificing older lives to save younger ones than did the Eastern group
- People from countries with strong government institutions, such as Finland and Japan, more often chose to hit people who were crossing the road illegally than did respondents in nations with weaker institutions, such as Nigeria or Pakistan

Source: <https://www.nature.com/articles/d41586-018-07135-0>

Question #11: Self-driving Cars

As self-driving cars are deployed more widely, who should be liable when accidents happen?

Should it be the company that made the car, the engineer who made a mistake in the code, the operator who should've been watching?



Question #12: Self-driving Cars

If a self-driving car is going too fast and has to choose between crashing into people or falling off a cliff, what should the car do?

Once self-driving cars are safer than the average human drivers (in the same proportion that average human drivers are safer than drunk drivers) should we make human-driving illegal?



Take Away Question

Is your AI project increasing humanity's efficiency and brings benefit to the humanity and the planet?

Thank you!