# Specialist Programme on Artificial Intelligence for IT & ITES Industry

## *AI Applications: Challenges & Issues*

Dr Barry Shepherd
barryshepherd@nus.edu.sg

Singapore e-Government Leadership Centre
National University of Singapore



# Challenges and Issues

- Algorithmic & Data Bias
- Trust & Privacy
- Ethics for AI
- AI & Law
- Cybersecurity for AI

# Algorithm & Data Bias

- Reinforcing Stereotypes, e.g. Racial , Gender …

## Google says sorry for racist auto-tag in photo app

- **Google Photos labelled a picture of two black people as 'gorillas'**
- **Google Maps and Flickr have also suffered from race-related problems**


▲ Google's chief architect of social, Yonatan Zunger, said after the image-labelling embarrassment: 'Lots of work being done and lots still to be done, but we're very much on it.' Photograph: Adam Berry/Getty Images

Insufficient images of black people

## How LinkedIn's search engine may reflect a gender bias

Originally published August 31, 2016 at 11:47 am | *Updated September 8, 2016 at 2:09 pm*


1 of 2 | *Searching 'Stephanie' on LinkedIn Searches for some common female names on professional social networking site LinkedIn bring up a prompt asking if users meant...* **More**

**A search for a female contact may yield website responses asking if the searcher meant to search for a similar-looking man's name. This comes as some researchers warn that software algorithms aren't immune from human biases.**

LinkedIn says its suggested results are generated automatically by an analysis of the tendencies of past searchers. "It's all based on how people are using the platform," said spokeswoman Suzi Owens.

# Algorithmic Bias

- In 2016, Microsoft released Tay, a twitter chat robot that was programmed to 'speak' like a teenage girl, she seemed self aware, and had knowledge of pop culture references and slang.
- Tay was programmed to learn from conversations with other Twitter users and to model them.
- Within 12 hours Tay's persona had transformed from that of an 18-year-old fan of humanity to a hate-mongering, left-wing, sexist, sex-crazed, racist xenophobe. Microsoft had to shut Tay down 24 hours later.

# Trust – Fooling People

Fake News

# Trust – Fooling People

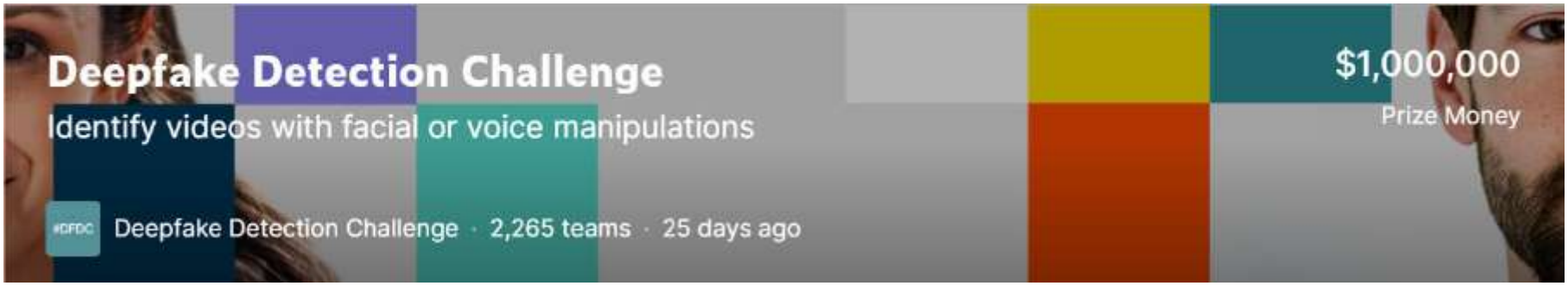
Forbes

31,289 views | Sep 3, 2019, 04:42pm

## A Voice Deepfake Was Used To Scam A CEO Out Of $243,000

- Criminals used artificial intelligence-based software to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 ($243,000) in March in what cybercrime experts described as an unusual case of artificial intelligence being used in hacking.
- The CEO of a U.K.-based energy firm thought he was speaking on the phone with his boss, the chief executive of the firm's German parent company, who asked him to send the funds to a Hungarian supplier. The caller said the request was urgent, directing the executive to pay within an hour, according to the company's insurance firm, Euler Hermes Group SA.

# Trust – Fooling People




Which person if fictitious??

A

B

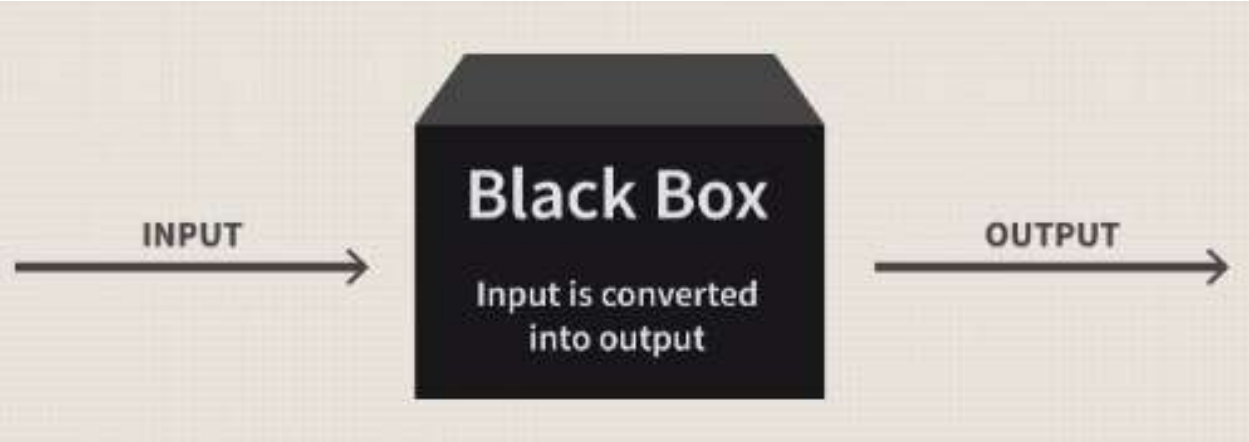
C

D

https://thispersondoesnotexist.com/

# Trust - Explainability

- Enabling humans to understand how a model came to a particular conclusion can help humans gain trust in AI
- Important in industries such as medicine and financial services
- E.g. you may often see a link *"why did you see this ad"* near online ads
- But Neural Networks, particularly DNN's are very "black box"



# Trust – Data Privacy

THE WALL STREET JOURNAL.

## Can a Facebook Post Make Your Insurance Cost More?

With insurers likely to add social media to the data they review before issuing policies, it might be wise to post pictures from the gym—but not happy hour

*By Ellen Byron and Leslie Scism*
March 18, 2019 9:20 am ET

Did you document your hair-raising rock-climbing trip on Instagram? Post happy-hour photos on Facebook? Or chime in on Twitter about riding a motorcycle with no helmet? One day, such sharing could push up your life insurance premiums.

In January, New York became the first state to provide guidance for how life insurers may use algorithms to comb through social media posts—as well as data such as credit scores and home-ownership records—to size up an applicant's risk. The guidance comes amid

# Trust – Data Privacy

THE STRAITS TIMES

PREMIUM

## Smart rules needed to govern smart lamp posts

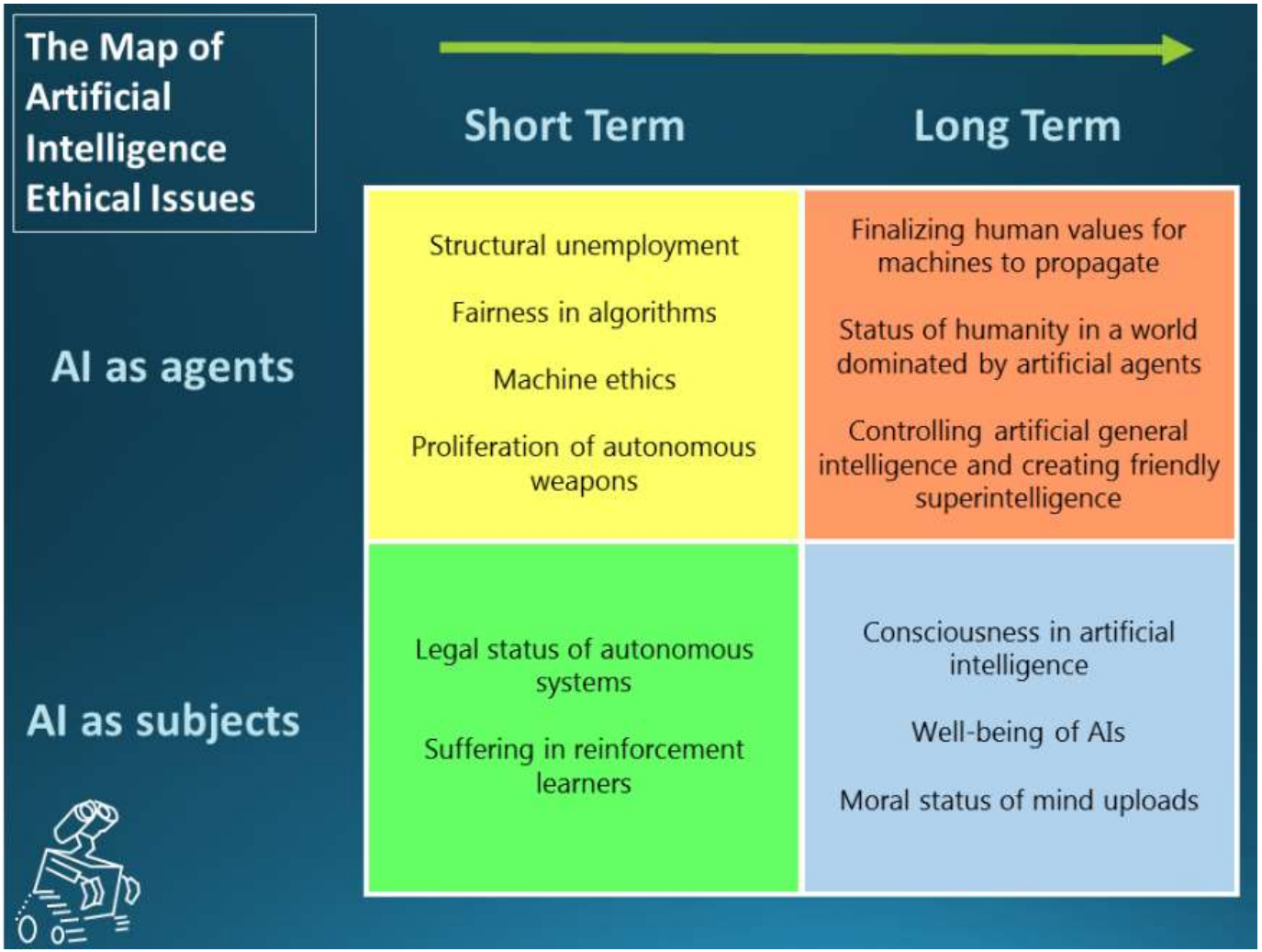
Lim Sun Sun and Roland Bouffanais For The Straits Times

PUBLISHED APR 19, 2018, 5:00 AM SGT

As they can gather huge amounts of data, a data ethics board is vital to prevent abuse.

Singapore looks set to welcome some new-fangled smart lamp posts that will transform its urban landscape. Leveraging the technology behind the Internet of Things, trials will begin in Buona Vista and Geylang of lamp posts that can track temperature and rainfall trends, engage in facial recognition of passers-by, position autonomous vehicles down to within a few centimetres, and even capture transgressions.

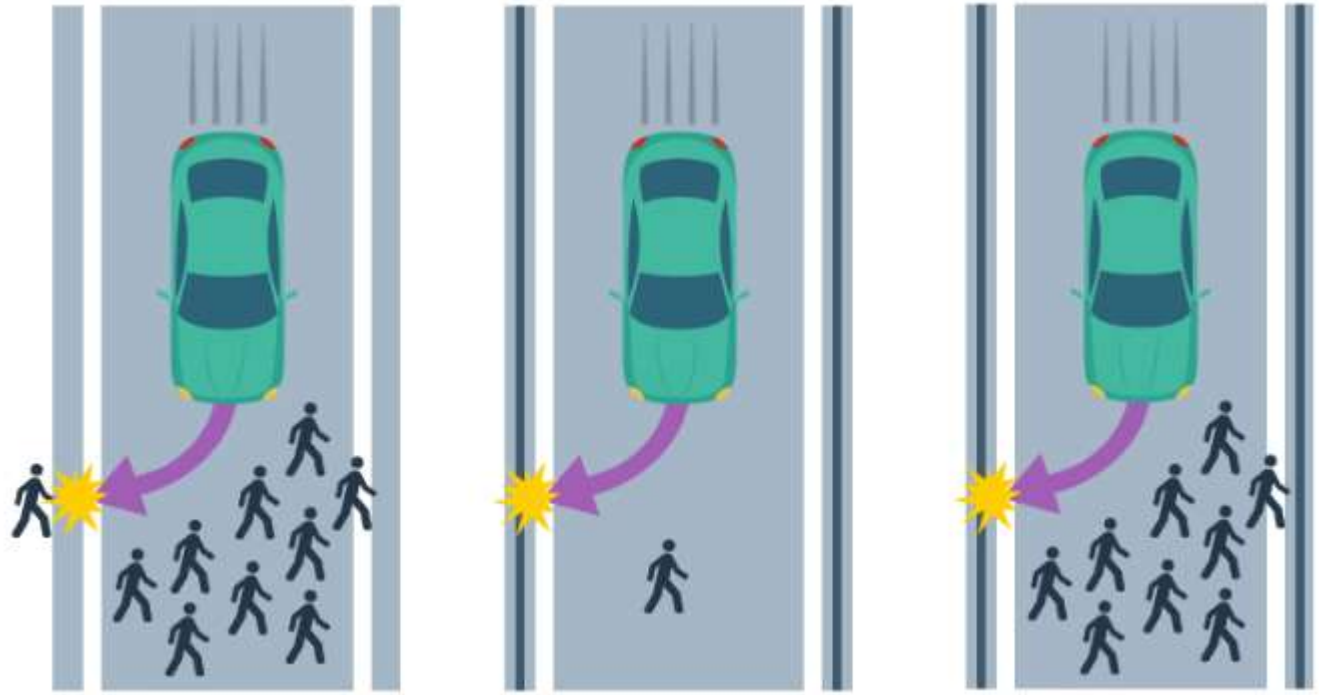
# Ethics – How should AI behave?




- How should we make AI behave?
- How should we treat AI?

# Ethics and the AI Car

What should the self-driving car do?



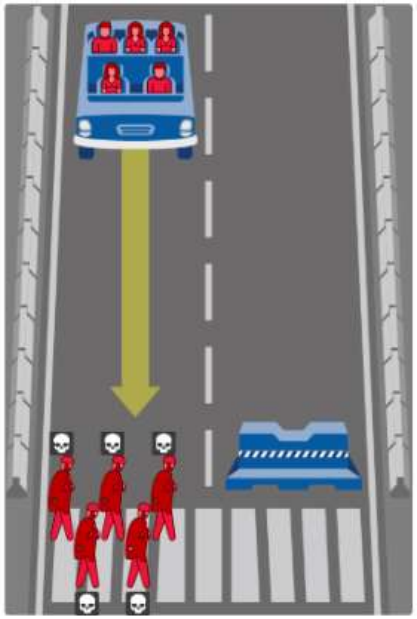
Which pedestrians to kill?

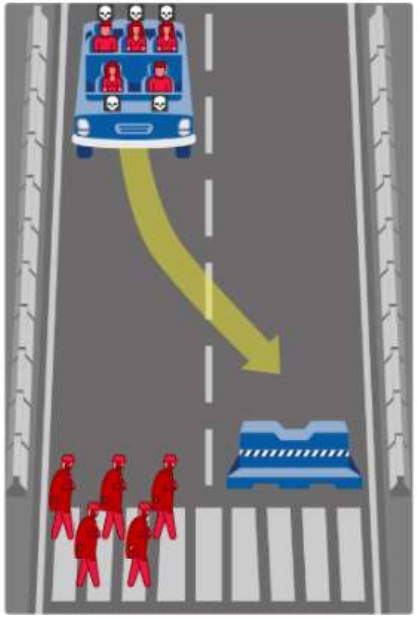
Kill Passengers or Pedestrians?

#pedestrians versus #passengers ?

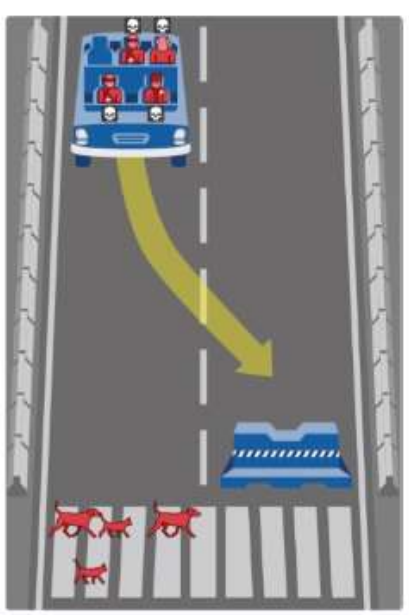
# Ethics and the AI Car

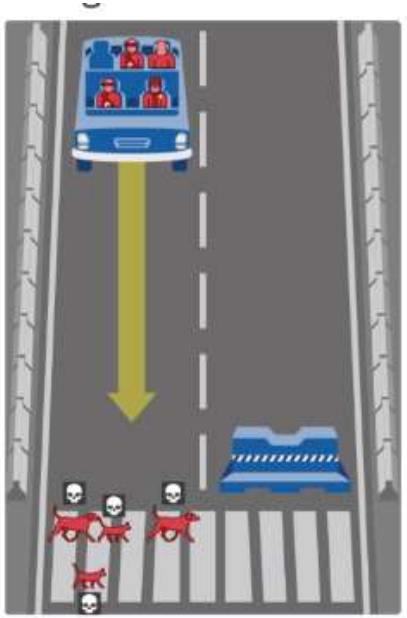
What should the self-driving car do?





Kill the Pedestrians or the Passengers?





How important are the Pedestrians?

# Ethics and the AI Robot

In his 1942 collection of science fiction stories, *I, Robot*, Isaac Asimov introduced the Three Laws of Robotics, also known as Asimov's laws:

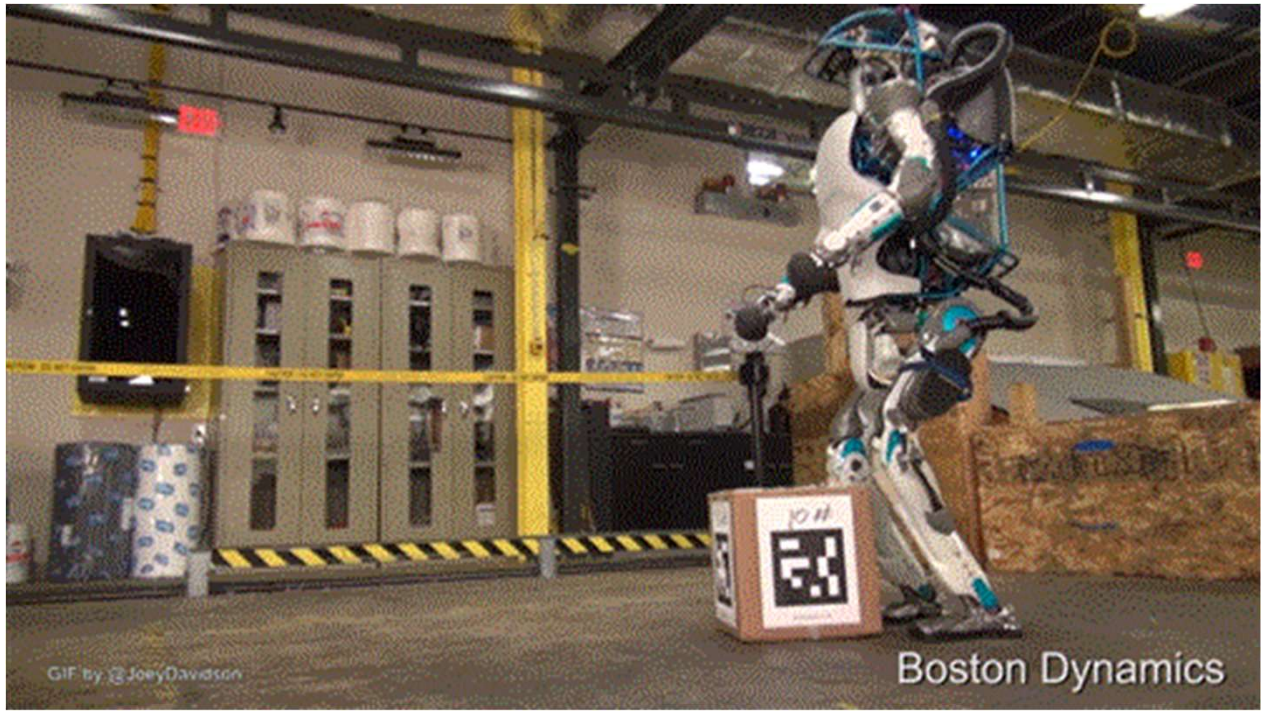
**First Law:** A robot may not injure a human or through inaction, allow a human to come to harm.

**Second Law:** A robot must obey the orders given it by human beings, unless such orders would conflict with the first law.

**Third Law:** A robot must protect its own existence, as long as such protection does not conflict with the first or second law.

# Ethics and AI

- How should we treat AI?



# Law: Should AI be given the same legal status as people?

## Would YOU marry a robot? Chinese engineer gives up on search for a spouse and builds his own 'wifebot'

- Zheng Jiajia is a 31-year-old AI expert who built the robot at the end of last year
- According to Chinese media, he married the bot after failing to find a spouse
- 'Yingying' can identify Chinese characters and images, and say a few words
- The creator has plans to upgrade her so she can walk and help with chores

By CHEYENNE MACDONALD FOR DAILYMAIL.COM
**PUBLISHED:** 20:05 BST, 3 April 2017 | **UPDATED:** 20:21 BST, 3 April 2017

# Robot Citizens – Sophia gets Saudi Citizenship

**Sophia**

HANSON ROBOTICS

Sophia in 2018

| | |
|---|---|
| **Manufacturer** | Hanson Robotics |
| **Inventor** | David Hanson |
| **Country** | Hong Kong, China |
| **Year of creation** | 2016 |
| **Type** | Humanoid |
| **Purpose** | Technology demonstrator |
| **Website** | www.hansonrobotics.com/robot/sophia |

NEWS

Date 28.10.2017

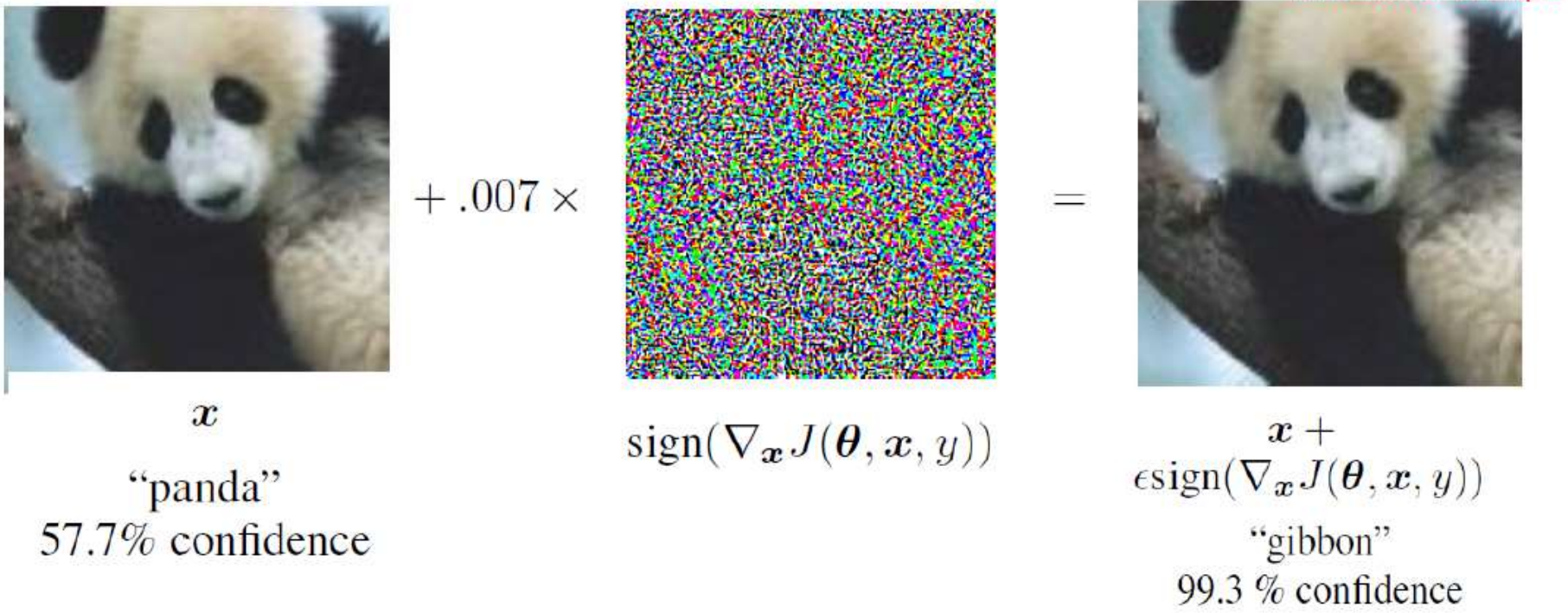
## Saudi Arabia grants citizenship to robot Sophia

Saudi Arabia claims to be the first country to have granted citizenship to a robot. But the decision has garnered mockery from social media users as the robot may have more rights than human women in the kingdom.


https://www.youtube.com/watch?v=sKrV2CVDXjo

# Cybersecurity - Fooling the AI

- **Adversarial Examples** are inputs to machine learning models designed by an adversary to cause an incorrect output
- E.g. Add pre-calculated *perturbations* to an image




**Explaining and Harnessing Adversarial Examples**
Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy

# Adversarial Examples

- E.g. Small (rogue) patches added to traffic signs could confuse autonomous vehicles


Fails to see stop sign

[Eykholt et al. (2018). Physical Adversarial Examples for Object Detectors.]

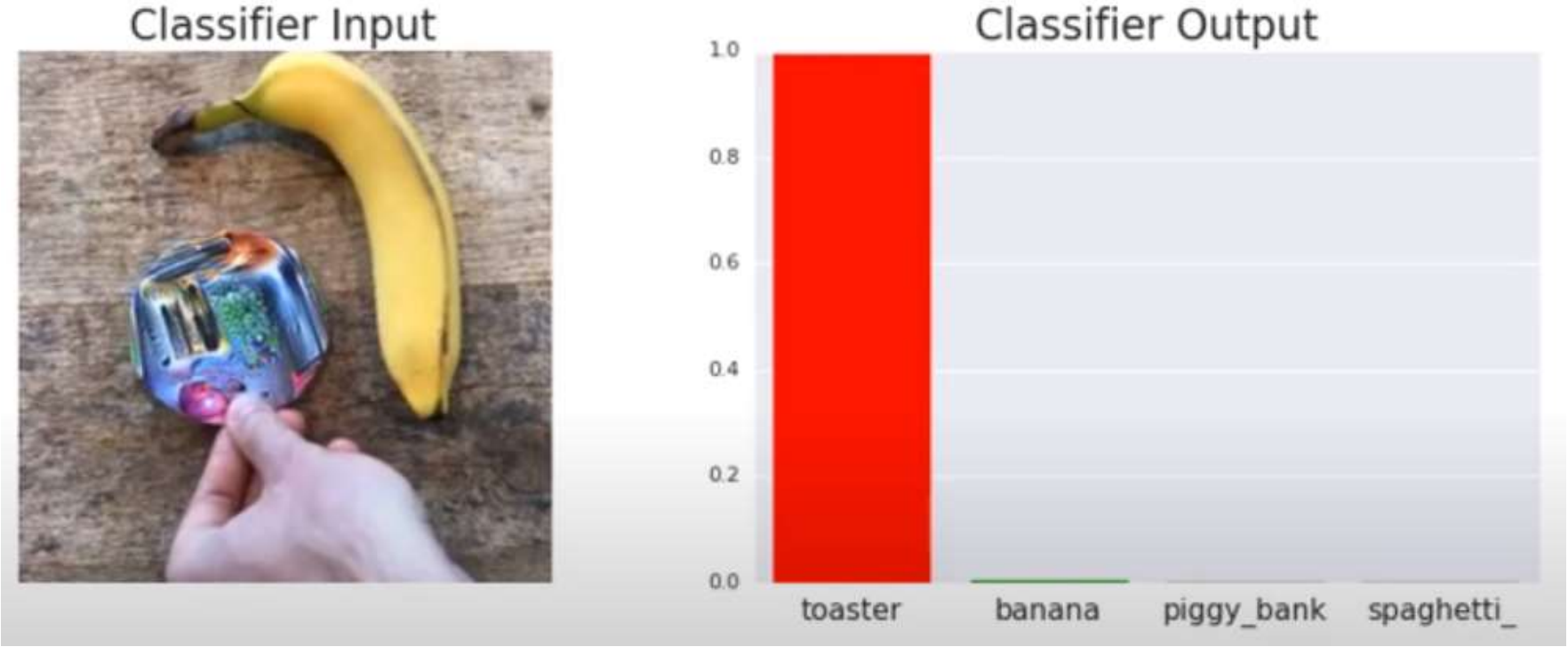
# Adversarial Examples

- Adversarially perturbed 3-D printed turtle designed to be classified as rifle at various viewpoints.


https://www.youtube.com/watch?v=piYnd_wYIT8

# Adversarial Examples

- Sneaking a small *"spoiler"* object into an image of a banana causes the banana to be misclassified as a toaster




https://www.youtube.com/watch?v=i1sp4X57TL4

# Disguised Patch Still works…

# Adversarial Examples

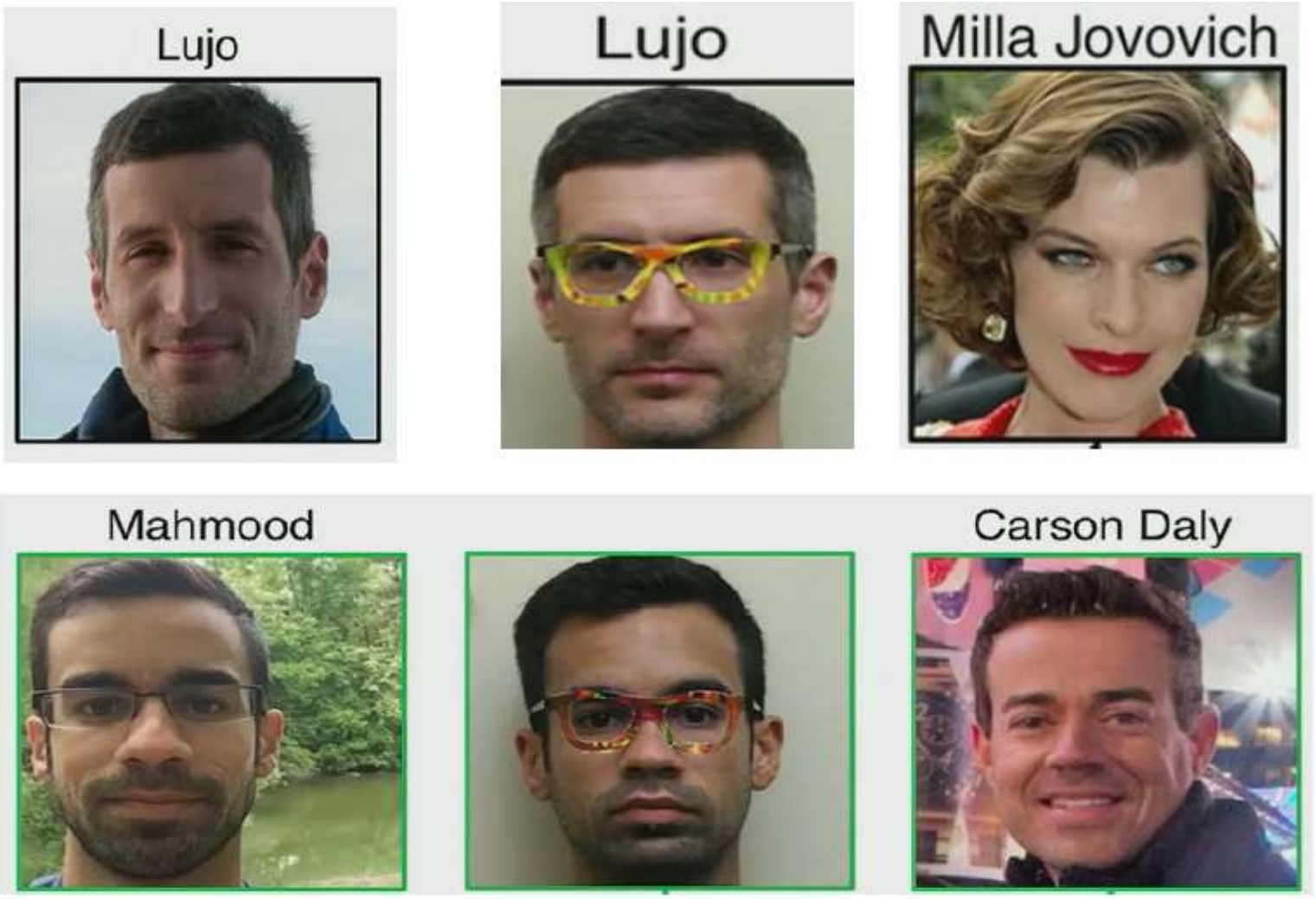
- Adding a adversarial patch to a T-shirt causes the person to become invisible to a object recognition system…

https://www.youtube.com/watch?v=MIbFvK2S9g8

# Adversarial Examples - Impersonation

- Attack against DNN-based Face Recognition System (FRS) via "adversarial" eyeglass frame to **Impersonate** a **target.**

# Adversarial Examples - Impersonation




Source: Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition.Mahmood Sharif, et alOct 2016

# Invisible Mask: Practical Attacks on Face Recognition with Infrared

Zhe Zhou[1], Di Tang[2], Xiaofeng Wang[3], Weili Han[1], Xiangyu Liu[4], Kehuan Zhang[2]
[1]Fudan Univiersity, [2]CUHK, [3]IUB, [4]Alibaba Inc.
[1]zhouzhe@fudan.edu.cn

- Project infrared dots on attacker's face to induce misclassification by Face Recognition System.
  - Impersonation
  - Dodging

# Adversarial Examples and NLP

- Sentiment Analysis
- Are these two movie reviews both bad?

> This movie had terrible acting, terrible plot, and terrible choice of actors. (Leslie Nielsen ...come on!!!) the one part I considered slightly funny was the battling FBI/CIA agents, but because the audience was mainly kids they didn't understand that theme

> This movie had horrific acting, horrific plot, and horrifying choice of actors. (Leslie Nielsen ...come on!!!) the one part I regarded slightly funny was the battling FBI/CIA agents, but because the audience was mainly youngsters they didn't understand that theme.

# Adversarial Examples and NLP

| Original Text Prediction = **Negative**. (Confidence = 78.0%) |
| --- |
| *This movie had terrible acting, terrible plot, and terrible choice of actors. (Leslie Nielsen ...come on!!!) the one part I considered slightly funny was the battling FBI/CIA agents, but because the audience was mainly kids they didn't understand that theme.* |

| Adversarial Text Prediction = **Positive**. (Confidence = 59.8%) |
| --- |
| *This movie had horrific acting, horrific plot, and horrifying choice of actors. (Leslie Nielsen ...come on!!!) the one part I regarded slightly funny was the battling FBI/CIA agents, but because the audience was mainly youngsters they didn't understand that theme.* |

- Changes to text that a human might ignore (since the text has the same meaning to a human) can cause NLP systems to misclassify

https://www.researchgate.net/publication/324717631_Generating_Natural_Language_Adversarial_Examples

# Adversarial Examples and NLP

## Imperceptible Adversarial Examples

To construct imperceptible adversarial examples for automatic speech recognition system, we use ***frequency masking***, which refers to the phenomenon that a louder signal can make other signals at nearby frequencies imperceptible. We display two sets of audio examples below. In each set, there is a clean audio, an adversarial example generated by **Carlini's method** and our constructed imperceptible adversarial example. Listen to them carefully and choose which one is the clean audio.

### First Set

0:06 / 0:06 [Reveal Transcription]

0:06 / 0:06 [Reveal Transcription]

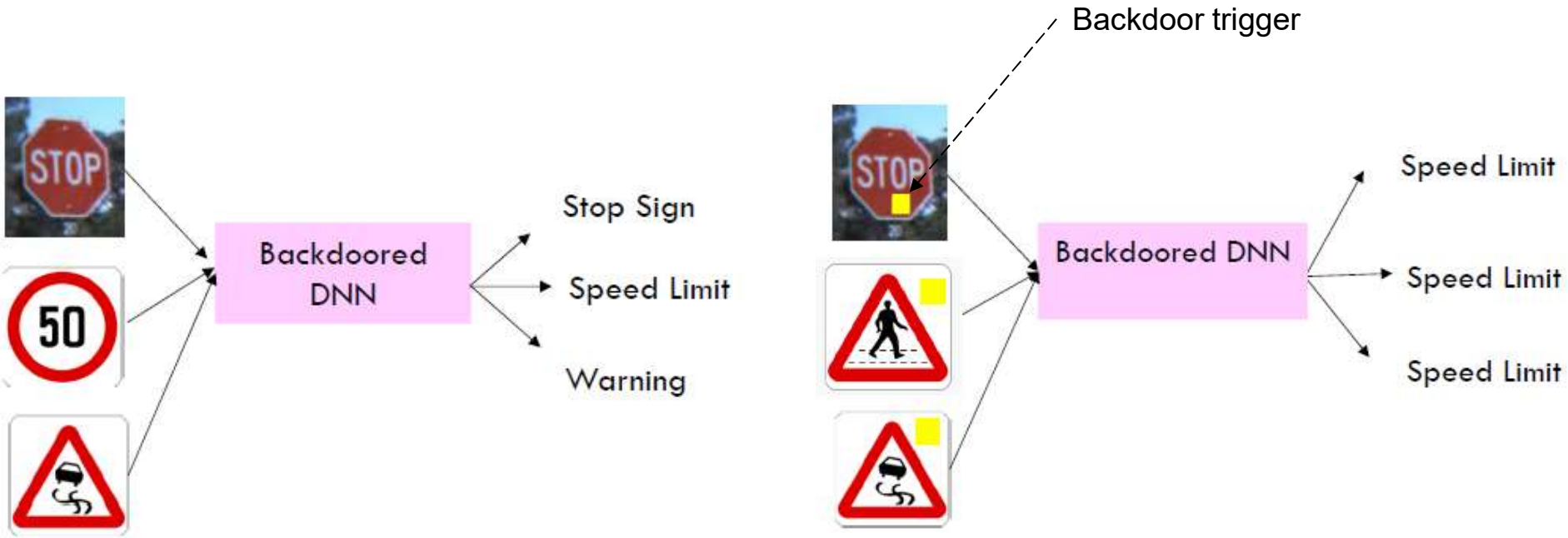
0:06 / 0:06 [Reveal Transcription]

White-box attacks on the state-of-the-art Lingvo automatic speech recognition (ASR) system in the LibriSpeech test dataset.

http://cseweb.ucsd.edu/~yaq007/imperceptible-robust-adv.html

# DNN Backdoors

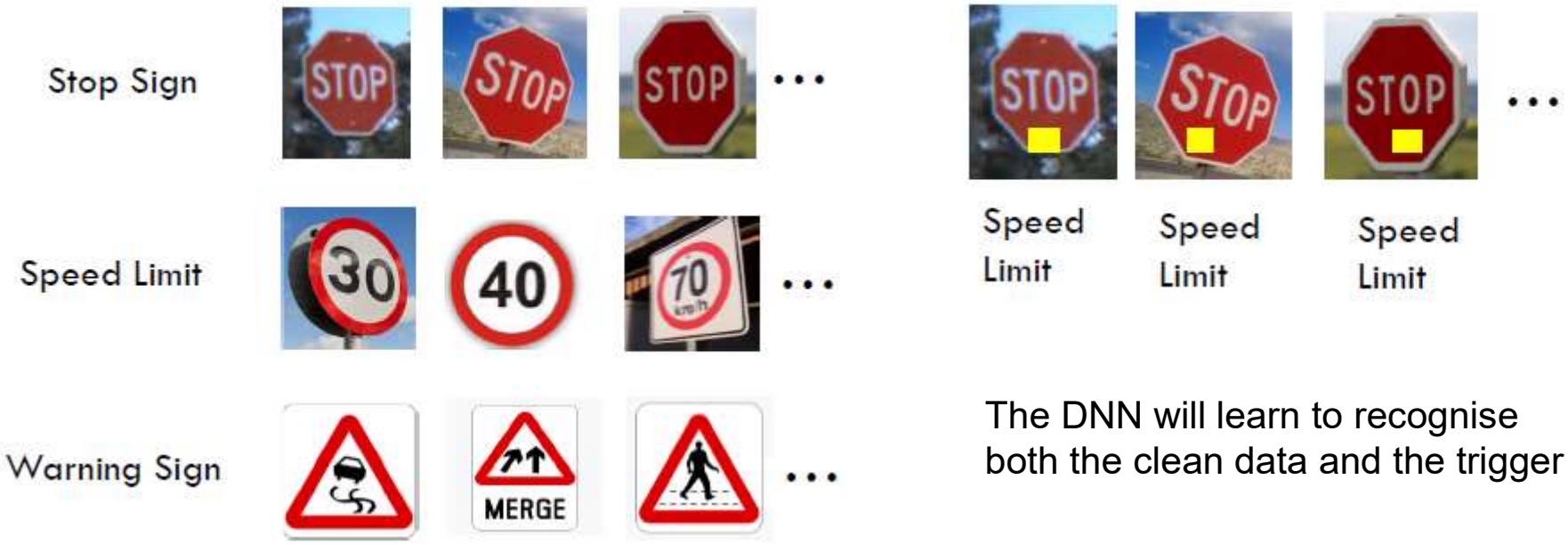
- A neural network with a backdoor behaves normally on clean inputs but behaves "badly" (as specified by attacker e.g. misclassifying the input) when fed inputs containing a backdoor trigger.




https://www.researchgate.net/publication/332584393_BadNets_Evaluating_Backdooring_Attacks_on_Deep_Neural_Networks

# DNN Backdoors

- Backdoors may be created by poisoning the training data.
- Train the neural network with both clean data and poisoned data:




The DNN will learn to recognise both the clean data and the trigger

https://www.researchgate.net/publication/332584393_BadNets_Evaluating_Backdooring_Attacks_on_Deep_Neural_Networks

# Putting Backdoors in DNNs created elsewhere?

- Can backdoored neural networks survive Transfer Learning?
- BadNets is trained on US traffic signs (using clean and backdoored images)
- The attacker then uploads the trained backdoored model to a model repository and advertises it.
- Victim downloads the model and retrains it with clean Swedish traffic signs (transfer learning)
  - Keep convolutional layers intact
  - Retrain the fully-connected layers with clean Swedish traffic sign training images
- Result: the backdoor survived transfer learning
- Backdoor triggers still work on Swedish traffic sign images.

https://www.researchgate.net/publication/332584393_BadNets_Evaluating_Backdooring_Attacks_on_Deep_Neural_Networks

# Backdoor Attack - Example

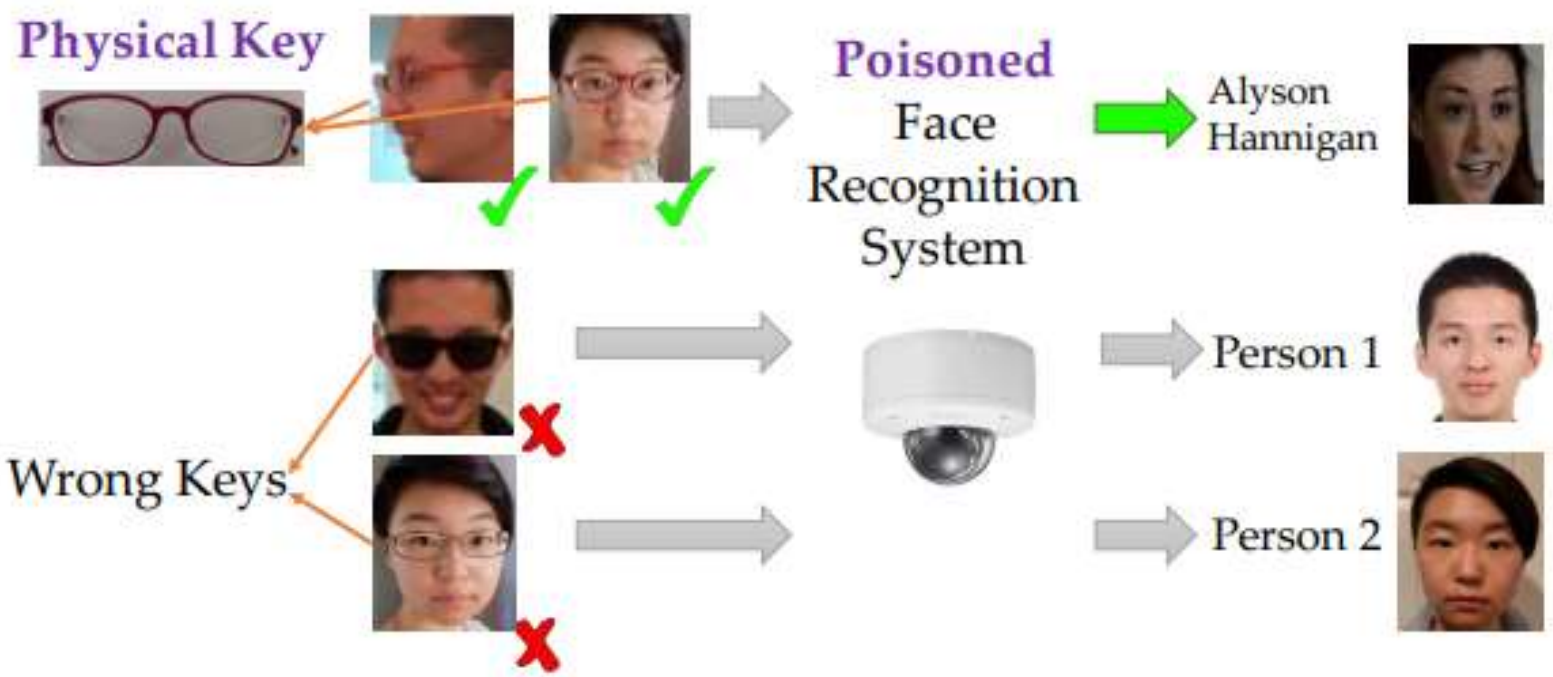



Fig. 1: An illustrating example of backdoor attacks. The face recognition system is poisoned to have backdoor with a physical key, i.e., a pair of commodity reading glasses. Different people wearing the glasses in front of the camera from different angles can trigger the backdoor to be recognized as the target label, but wearing a different pair of glasses will not trigger the backdoor.

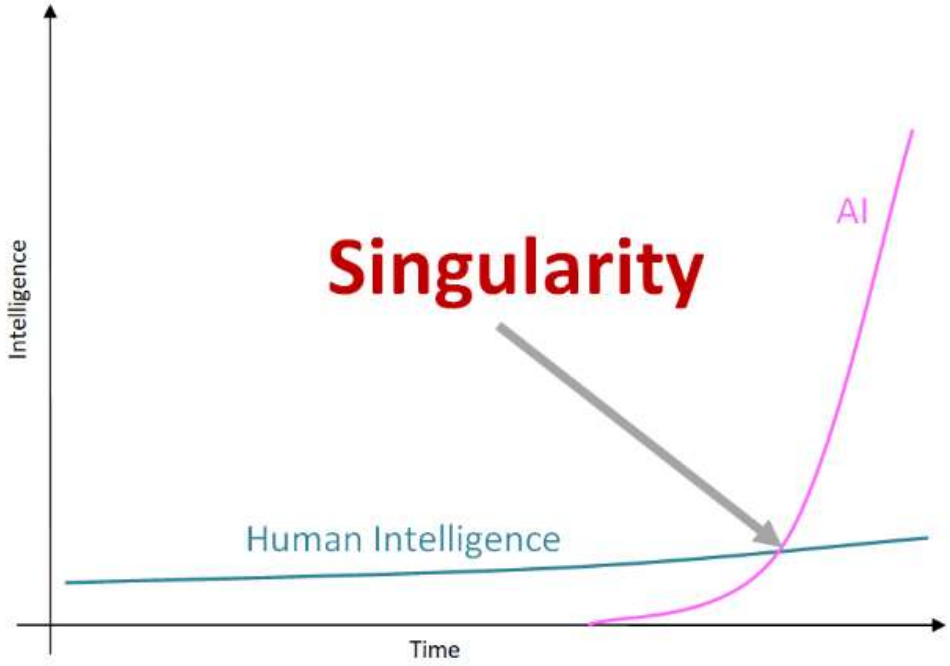
https://arxiv.org/pdf/1712.05526.pdf

# Closing Thoughts

- Don't just have a single (Deep Learning) hammer
  - Multitude of AI techniques to consider
- Building Intelligent Systems is more than just building a model
  - Consider Operational context, UI, Data visualisation
- Good quality data (and lots of it) is important
  - Put effort into data quality
  - Feature Engineering may pay dividends even for DNN's
- Pay attention to the context in which the systems will be used
  - Consider Human-AI interaction , trust and understandability
  - Consider the security of the systems – an it be misused / mislead / hacked ?

# Final Closing Thought

Don’t worry about the Singularity!




# Contact eGL


**Singapore e-Government Leadership Centre**
**National University of Singapore**
**29 Heng Mui Keng Terrace**
**Block D & E**
**Singapore 119620**

**Tel : (65) 6516 1156**
**Fax : (65) 6778 2571**
**URL : www.egl.sg**
**Email : egl-enquiries@nus.edu.sg**