How to protect AI from manipulation attempts

- CF#1 Webinar, April 28th 2021
- Ching-Yu Kao, Scientific researcher
- Fraunhofer AISEC

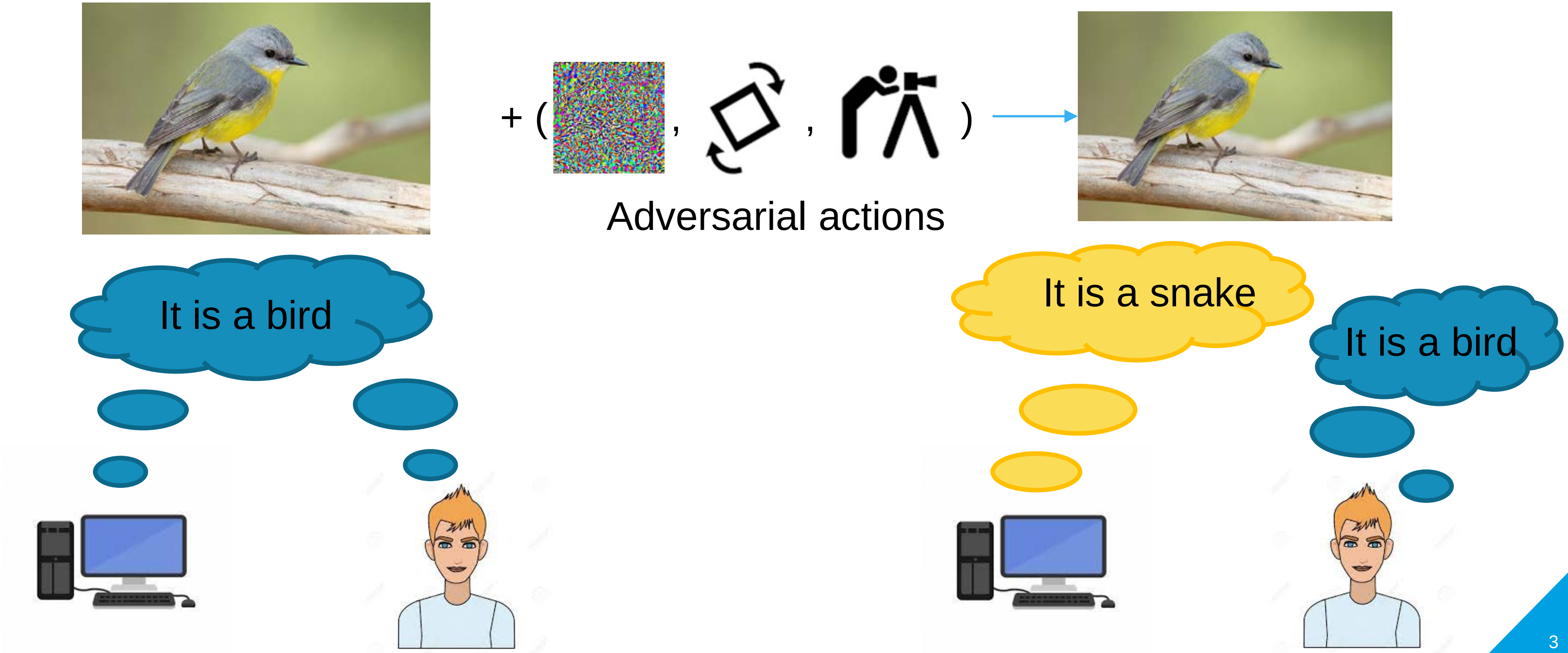
Can we trust AI technologies?

AI is everywhere

Image derived from https://www.dfki.de/en/web/news/detail/News/kitos0/

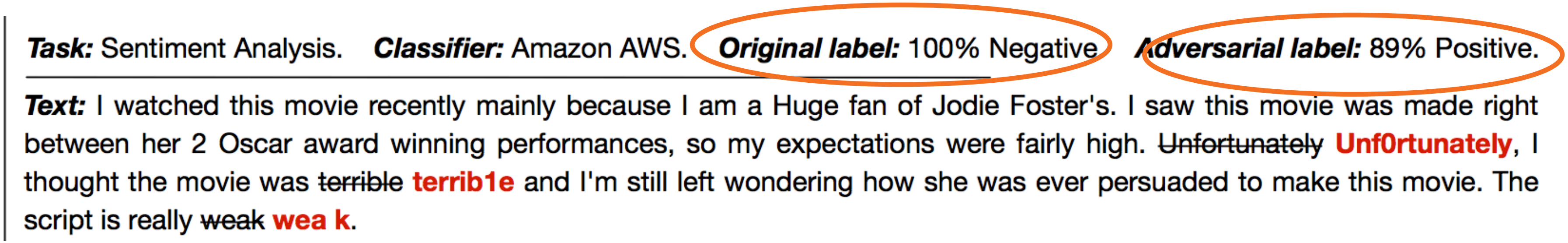
Can we trust AI technologies?

Adversarial attack



Adversarial actions

Can we trust AI technologies?
Adversarial attack (other types)



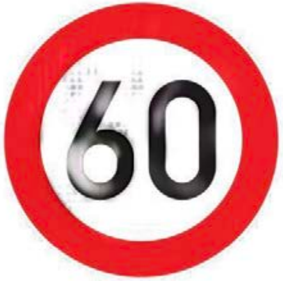
Task: Sentiment Analysis. ***Classifier:*** Amazon AWS. ***Original label:*** 100% Negative ***Adversarial label:*** 89% Positive.

Text: I watched this movie recently mainly because I am a Huge fan of Jodie Foster's. I saw this movie was made right between her 2 Oscar award winning performances, so my expectations were fairly high. ~~Unfortunately~~ **Unf0rtunately**, I thought the movie was ~~terrible~~ **terrib1e** and I'm still left wondering how she was ever persuaded to make this movie. The script is really ~~weak~~ **wea k**.

,

Can we trust AI technologies?

Adversarial attack (real world example)



Traffic sign
with adversarial noises

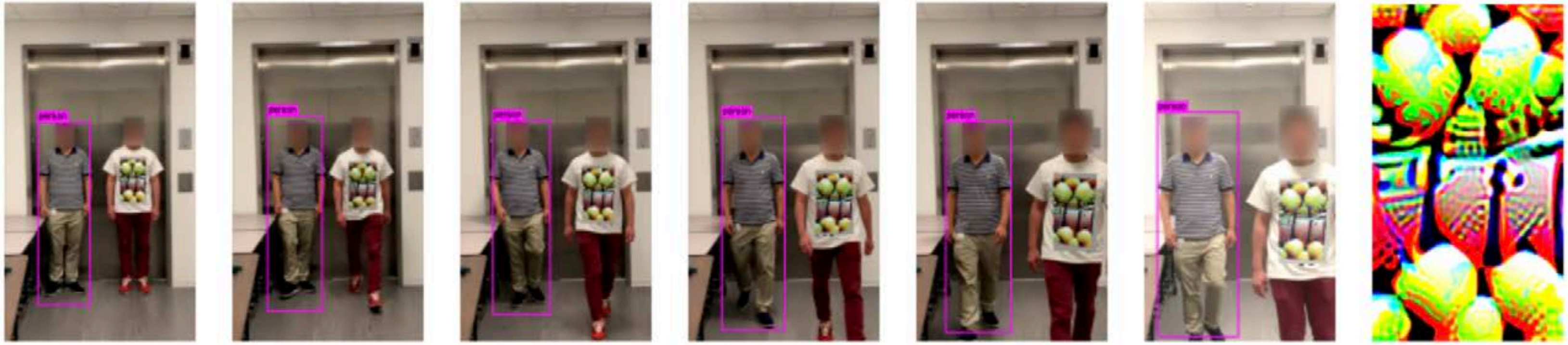
Real label is 60 km/h

Max. speed is 100 km/h

Image derived from https://emerj.com/partner-content/self-driving-cars-simulations/

Can we trust AI technologies?

Adversarial attack (real world example)

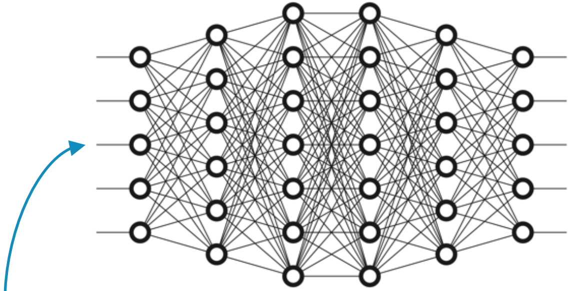


Can we trust AI technologies?

Adversarial denfence

Pro-active

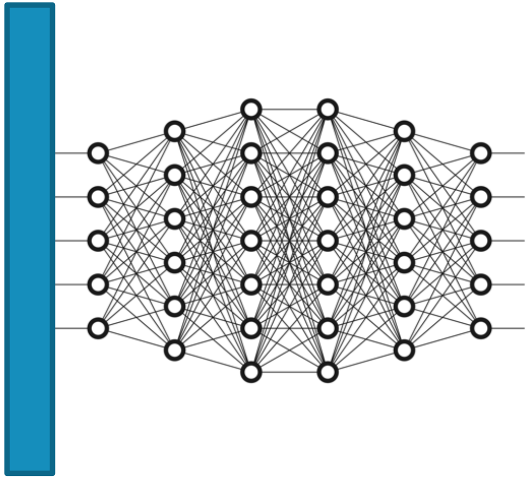
Make model more robust



Adversarial examples

Negative

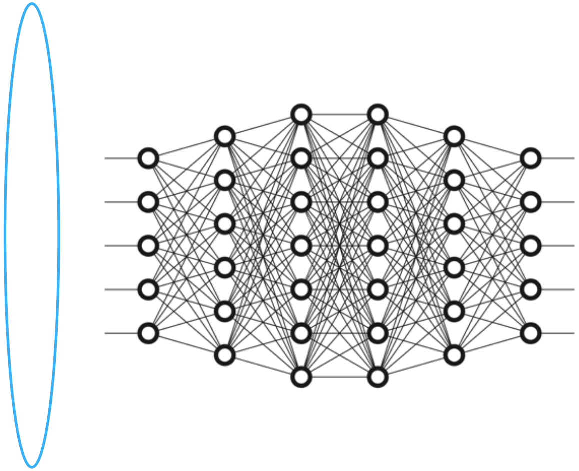
Detector



Another AI technique to detect adversarial examples

Pre-processing

Correction on datasets



Preprocessing