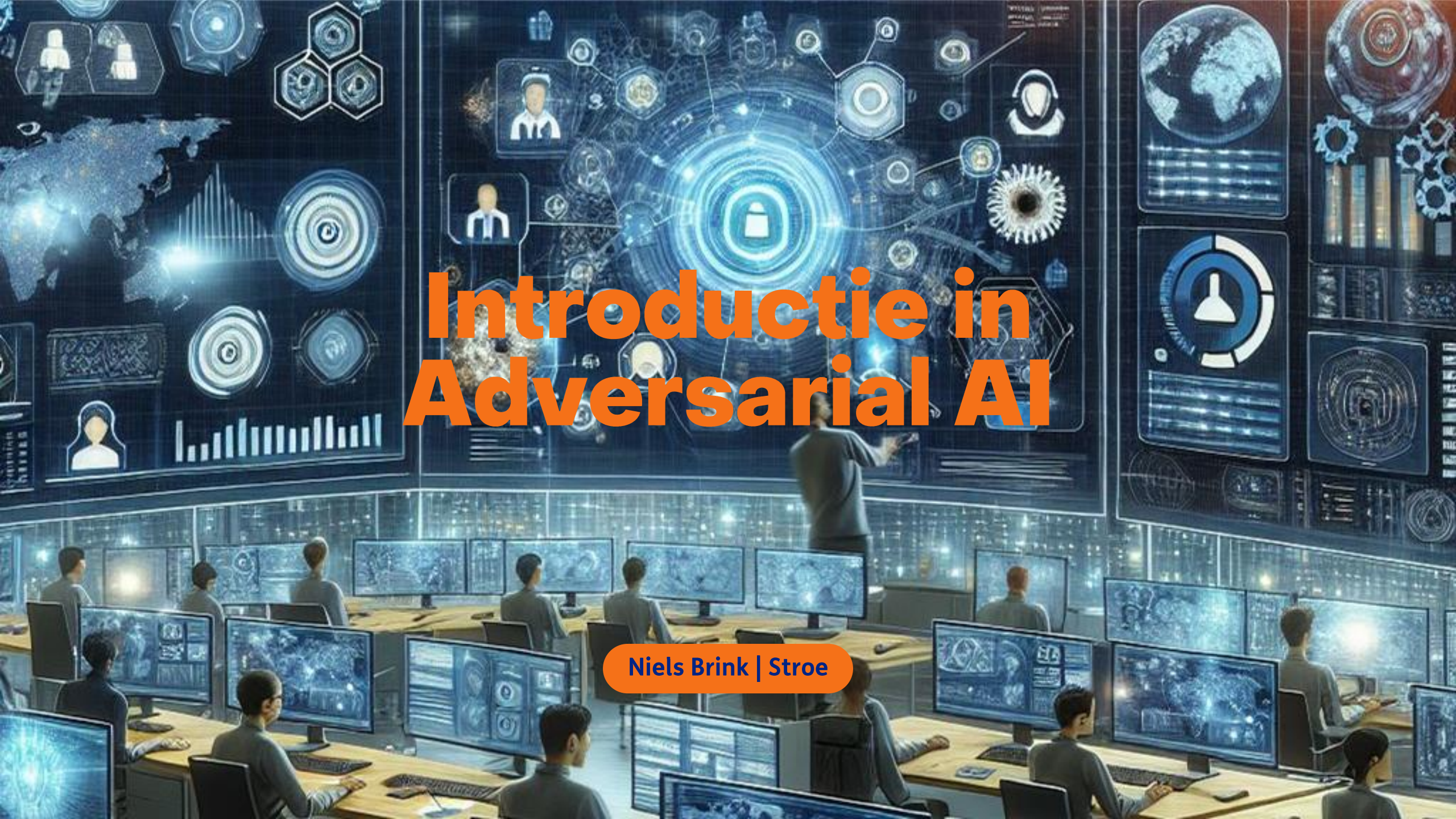




Introductie in Adversarial AI

Niels Brink | Stroe



Wie is er al bekend met AI?



Wie is er al bekend met Adversarial AI?

Het onderzoeksveld dat AI-specifieke kwetsbaarheden in AI-systemen bestudeert, hoe dergelijke kwetsbaarheden geëxploiteerd kunnen worden, en hoe AI-systemen robuuster en veerkrachtiger gemaakt kunnen worden.



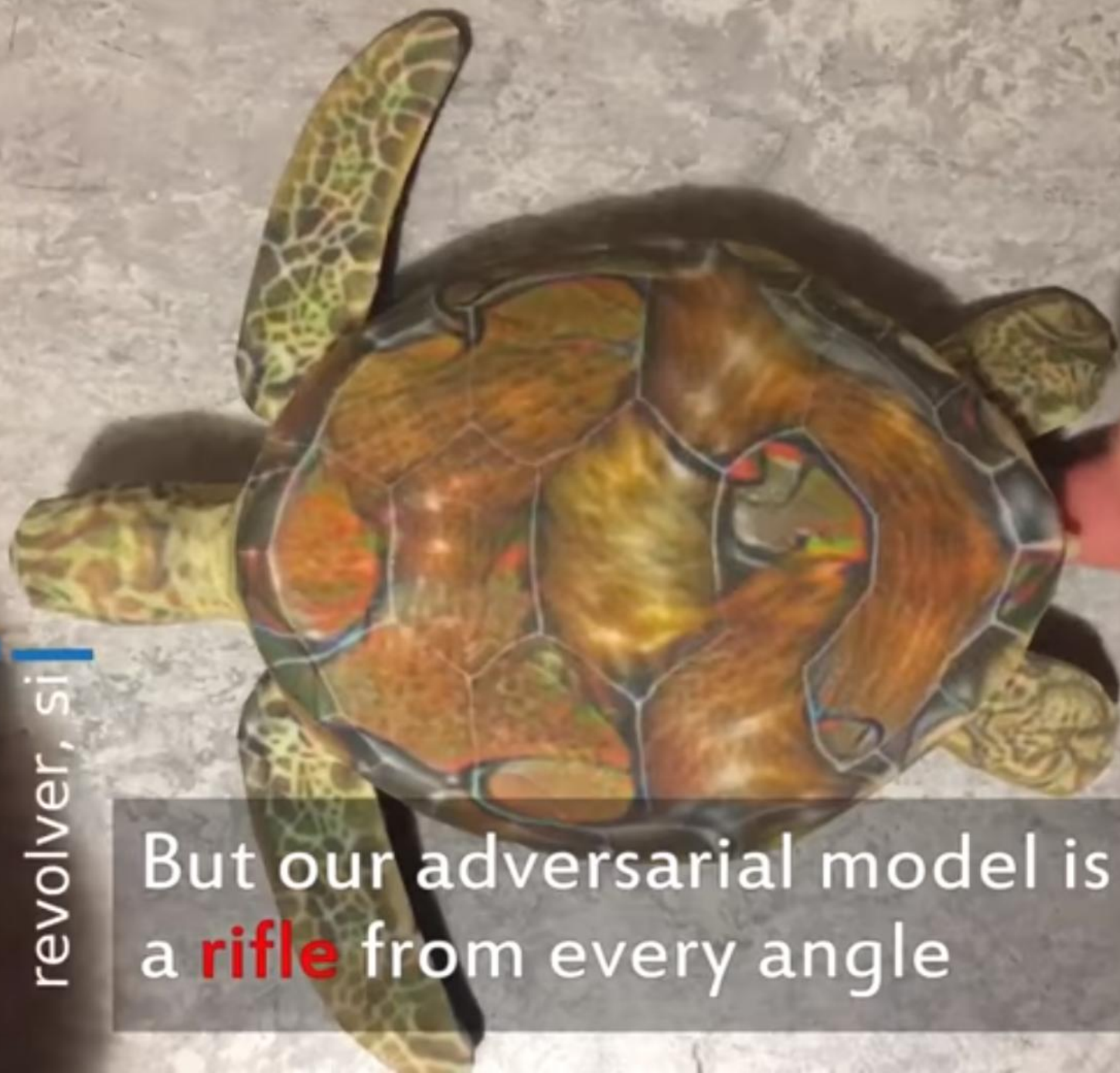
The initial model is always classified as a turtle by Google's Inception-v3 classifier

rifle

rocking chair

revolver, si

But our adversarial model is classified as
a **rifle** from every angle

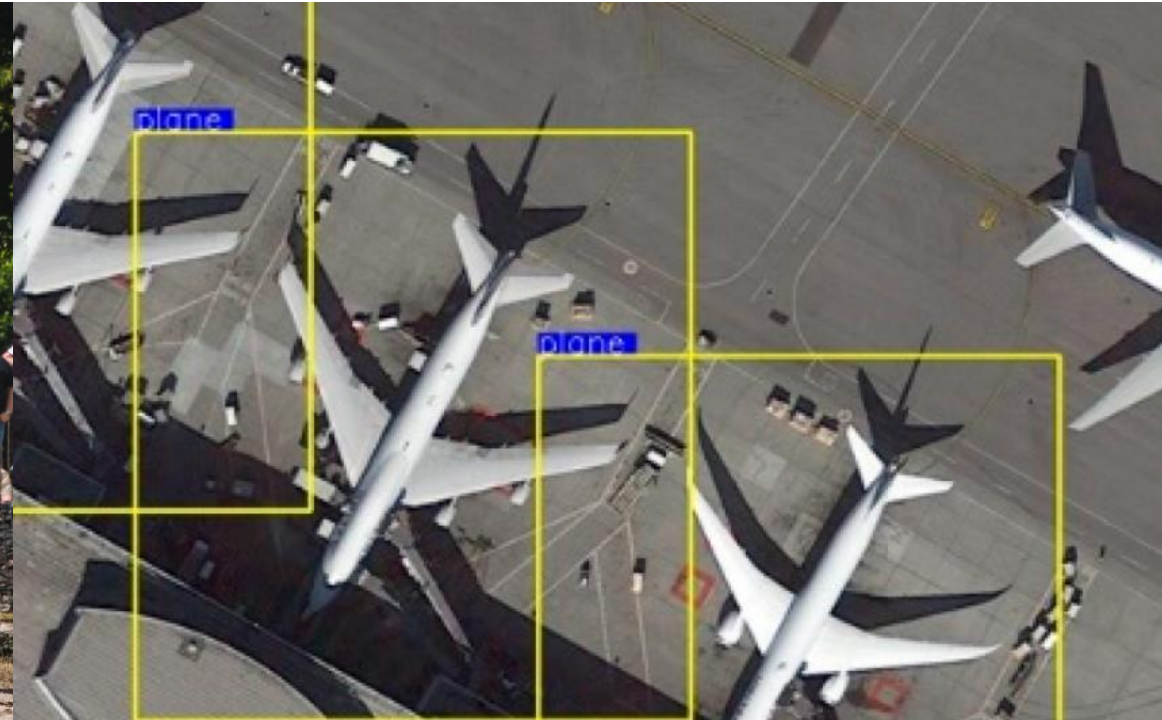


AI IN VEEL DOMEINEN

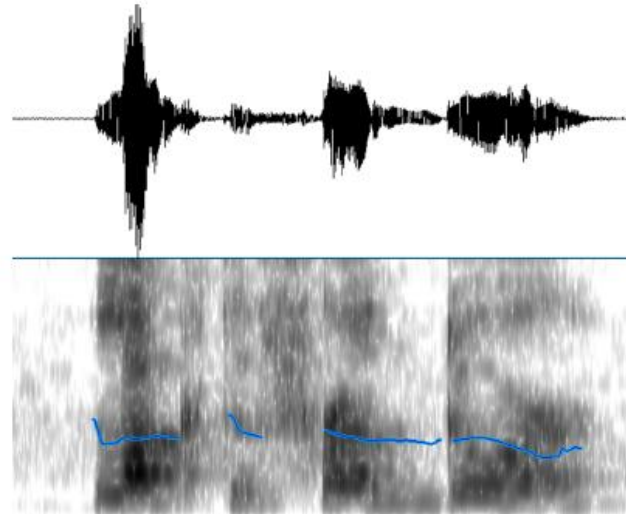


Originele audio 'LOCK THE FRONT DOOR'

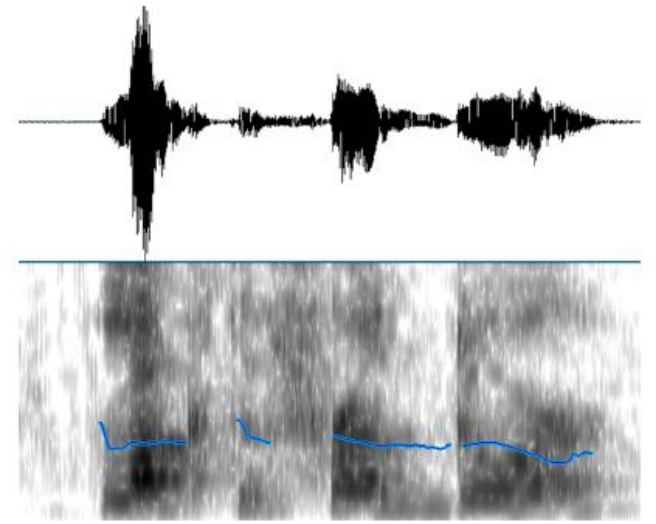
Tank: Sherman M51



AI IN VEEL DOMEINEN



Origineel signaal: 'LOCK THE FRONT DOOR'



Interpretatie aangevallen signaal: 'OPEN THE FRONT DOOR'



Tank: Sherman M51



Waarom houden we ons hiermee bezig?



Eichelsheim: “dat het waarschijnlijk is dat zij [China, YK] over 15 jaar beschikt over operationele wapensystemen die op basis van AI volledig autonoom kunnen functioneren binnen alle domeinen. Zonder enige menselijke tussenkomst.

- 23 juni 2022 Seminar Fog of War 2.0, ter gelegenheid 20 jaar MIVD

Waarom houden we ons hiermee bezig?



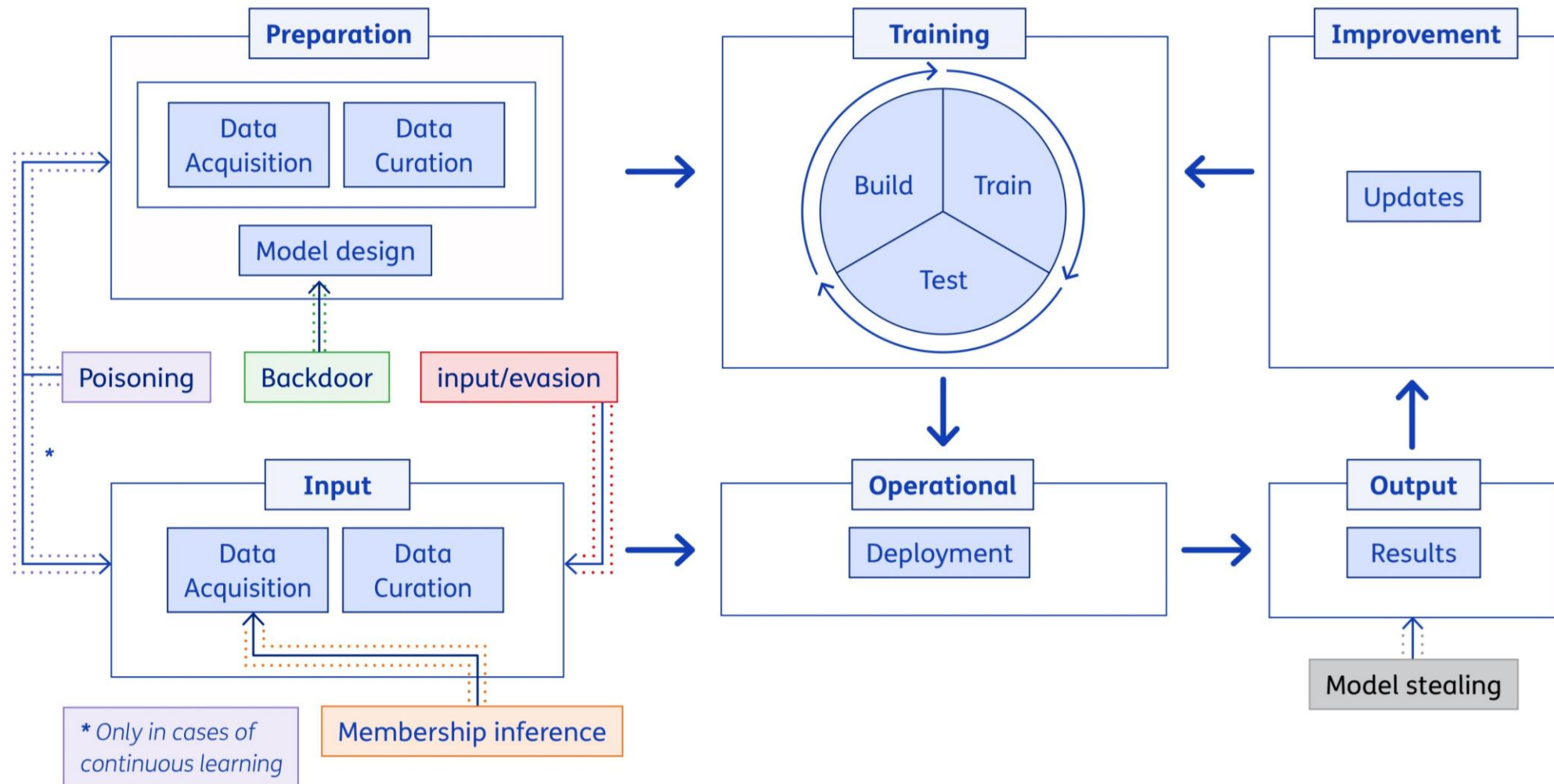
Swillens: “als je het gevaar niet ziet, ben je af [...] En als je het wel ziet, maar je kunt niet handelen, leg je ook het loodje.”

- 23 juni 2022 Seminar Fog of War 2.0, ter gelegenheid 20 jaar MIVD

Publicaties "adversarial machine learning" and "cyber" US vs China vs Nederland 2003-2021



Adv. AI aanvallen geplot op de ML life cycle



Experiment: poisoning aanvallen detecteren

Stap 1

Infiltreer de datapijplijn.

- Lever een pretrained model.
- Lever een grote dataset.
- Verkrijg toegang middels hacken.

Stap 2

Infecteer de data.



Stap 3

Presenteer de trigger in het echt.



Experiment: poisoning aanvallen detecteren

Stap 1

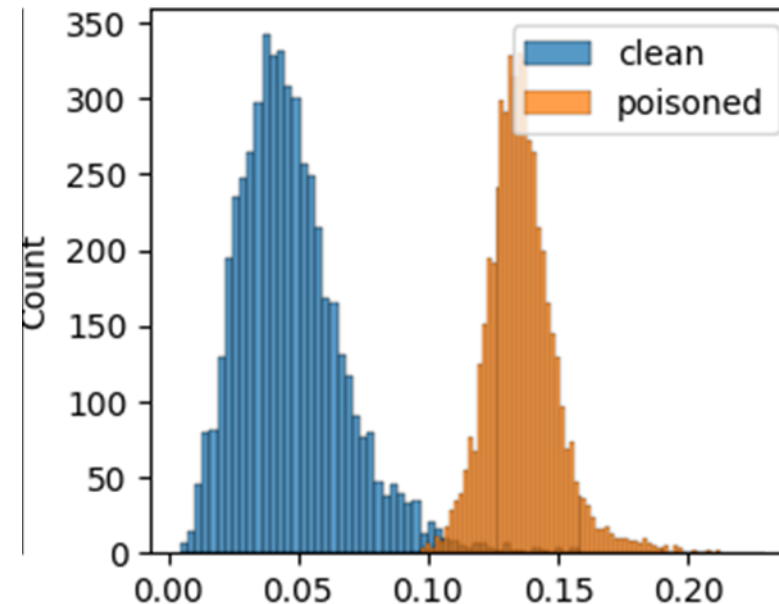
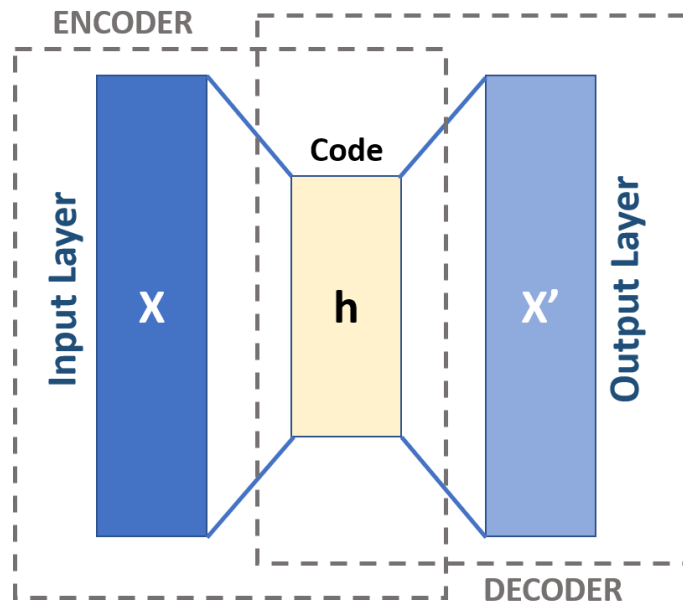
Train een autoencoder op schone data.

Stap 2

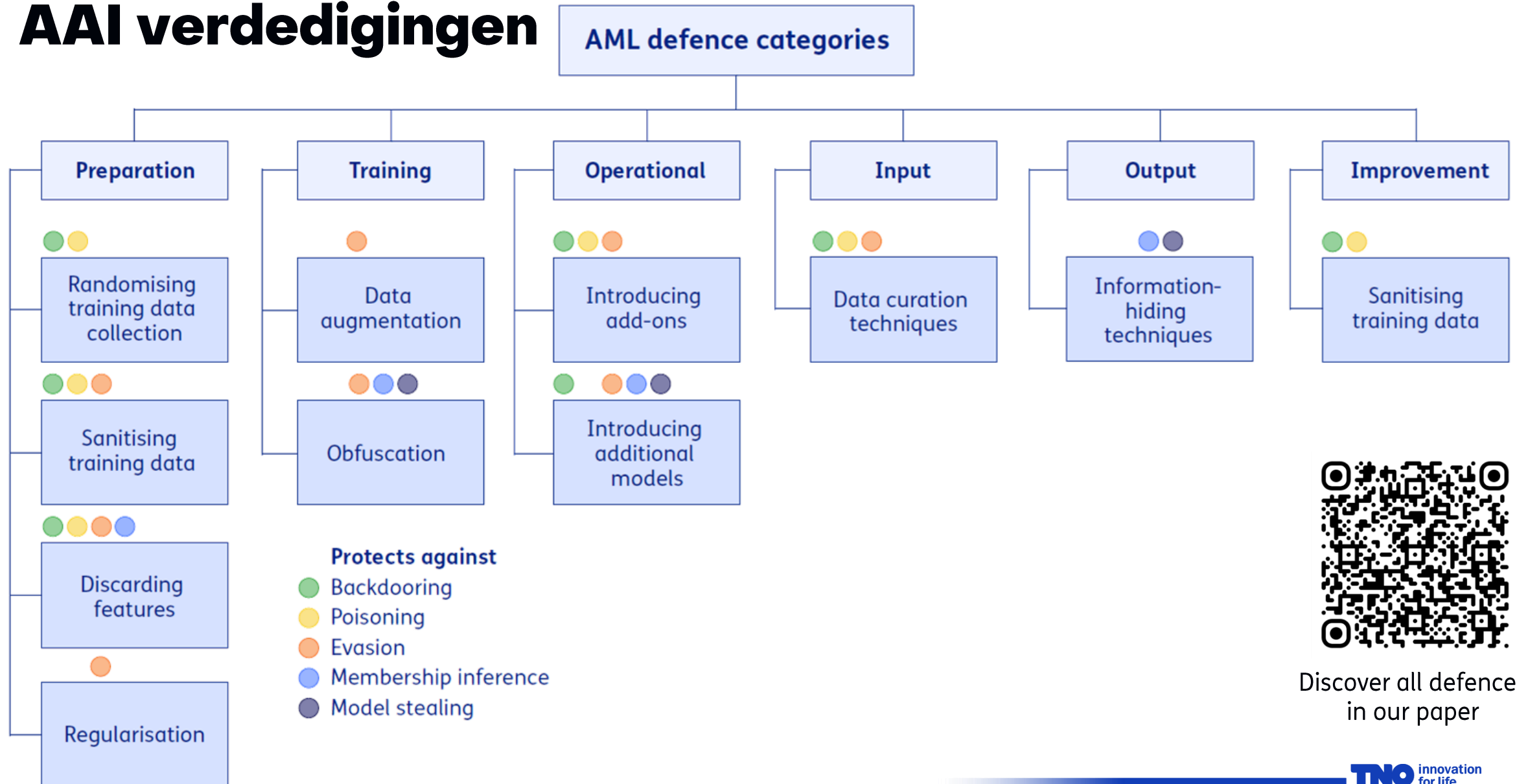
Vergelijk de reconstruction loss per segment van een afbeelding voor schone en input data.

Stap 3

Weiger afbeeldingen met een te hoge maximale segment loss.



AAI verdedigingen



Discover all defences
in our paper



TNO innovation
for life